1

Proceedings

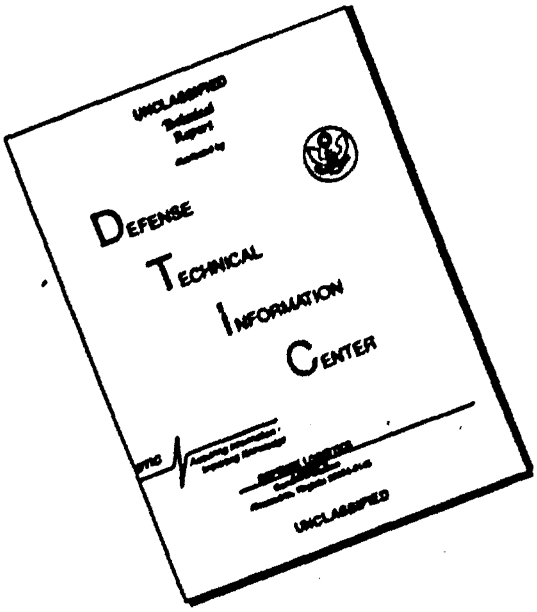
WORKSHOP ON RELIABILITY IN COMPUTATIONAL MECHANICS

October 26-28, 1989
Austin, Texas

J. T. Oden, Editor

To be published in a special issue of *Computer Methods in Applied Mechanics and Engineering*, to appear in September, 1990.

DISCLAIMER NOTICE



UNCLASSIFIED

DEFENSE TECHNICAL INFORMATION CENTER

DTIC

UNCLASSIFIED

THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

A

Preface

RELIABILITY IN COMPUTATIONAL MECHANICS

Much of what engineers and scientists do is to model natural phenomena. They develop mathematical models of nature so as to study and predict the behavior of physical systems. The remarkable advances in technology over the last half century attest to the success of this approach. Mathematical models do indeed "work." Their use represents a proven approach toward scientific discovery and engineering analyses and design, and one can safely predict that the confidence in results of mathematical modeling will grow as further proof and experience accumulates as to their utility and their reliability. Indeed, it is this latter quality, *reliability*, that emerges as the key to further progress in computational mechanics.

There has been growing concern about the issue of reliability in computational modeling in recent years. Success of computational modeling of certain classes of linear problems may have lulled many into a false sense of confidence in computed results. Exactly how reliable are contemporary computational modeling procedures? How can this reliability be assessed? What factors affect it? How can reliability be improved? Indeed, what directions must future research in computational modeling take to increase reliability of the more sophisticated models needed to simulate phenomena of importance in engineering?

These were the issues that led to the organization of the Workshop on Reliability in Computational Mechanics held at Lakeway, Austin, Texas in October, 1989. An international group of mathematicians, engineers, and scientists working in computational mechanics met for three days to discuss and study this subject. This volume contains invited papers and selected contributed papers presented at this meeting. The papers presented at the Workshop fell into four broad categories:

1) Mathematical modeling;
2) *A priori* analysis, including principles of convergence, robustness and their reliability;
3) *A posteriori* analysis, including adaptive methods; and
4) Computer aspects of modeling such as mesh generation, solid modeling and their reliability.

In addition, papers on parallel computing, applications to practical problems, selection of benchmark problems for code verification, and related issues were discussed. The majority of papers focused on finite element methods and their applications, but a number of papers also dealt with boundary element methods, finite difference methods, and spectral methods as well. (KR)

The order of the papers presented in the present volume essentially follows that of the order in which they were presented at the Workshop and this order was selected primarily to present a connected sequence of presentations covering the major topical areas. The papers by Bathe, Lee, and Bucalem; Desai, Wathugala, Sharma, and Woo; Babuska, Shephard, Baehmann, Georges, and Korngold; and Noor, Burton, and Peters dealt with issues of developing mathematical models of engineering analysis. Issues of convergence and *a priori* error estimation are dealt with in the papers by Brezzi and Bathe; Szabo; Cowsar, Dupont and Wheeler; and Arnold. *A posteriori* error estimation and adaptivity are discussed in the papers by Ewing, Zienkiewicz and Zhu; Oden, Demkowicz, Rachowicz and Westermann; Planck, Stein and Bischoff; Johnson; Bank and Welfert; and Wang and Carey. Applications and computational issues are taken up in the papers of Benantar, Biswas and Flaherty; Shephard, Baehmann, Georges, and Korngold, and to an extent were dealt with in part in a number of the other papers mentioned above.

The invited papers at this meeting were organized so as to present not only a tutorial style presentation of basic techniques in modeling, *a priori* and *a posteriori* error estimation, but also to deal with applications of these subjects to contemporary problems. Thus, the volume includes an interesting mixture of application of existing methods to issues of reliability as well as studies of new methods that touch upon or depend upon the reliability of computational techniques in mechanics.

At the end of the workshop, a number of principal conclusions were identified which generally reflect the results of the papers collected here. A summary of these is given as follows:

1. Reliability in computational mechanics is strongly affected by the choice of the mathematical model, and in most instances there is wide latitude in possible choices of models, boundary conditions, material properties, loads, initial conditions, etc. Minor differences in modeling features can create dramatic changes in results. The general problem of model design is classical in mechanics and is difficult to formulate in precise terms owing to the difficulty in specifying what an [illegible]able or optimal model is in specific applications. Very often, one model is acceptable for si[illegible] one feature of a problem (e.g. energy or stress-concentration factor) but this same model [illegible] inadequate for simulating other features.

Some progress in quantifying the modeling problem was reported in the Workshop in the context of hierarchical models of beams, plates, and shells in which the three-dimensional linear elasticity solution is viewed as the standard to which other models are compared. Similar modeling programs are under study for nonlinear materials wherein a given material response is presumed to be modeled by theories which form a part of a hierarchy of material models. Further work in this area is needed as the subject is fundamental to engineering analysis and design.

2. Once a mathematical model of a natural event is defined, the problem reduces to one of numerical analysis and computation. The reliability of the computational model has become an issue of increasing concern as computational models are used with increasing frequency in engineering designs. The use of *a posteriori* error estimates has become an accepted means for assessing and controlling computational error.

3. Theories and methods for *a posteriori* error estimation are available for certain classes of linear elliptic problems. Additional work on the mathematical theory supporting these estimates is needed, but significant progress has been made in the last three to five years. Many of these methods are being used in other classes of problems with mixed success. Analyses of error estimates for nonlinear and time dependent problems is needed. Some progress in these areas was evident at the Workshop, including estimates of errors in Euler and Navier Stokes equations in computational fluid dynamics and in error estimation for parabolic and hyperbolic partial differential equations.

4. *A priori* error estimation continues to be important in assessing the convergence properties of various numerical techniques. Equally important, *a priori* error analysis provides a basis for the design of acceptable numerical schemes and for the comparison of one scheme with another. However, a new practical use of existing *a priori* estimates is used as a basis for adaptive h- and p-refinement. Several techniques for adaptive finite element methods were presented in the Workshop that effectively used *a priori* estimates as a basis for error control and mesh adaptation.

5. Many of the concepts and strategies of adaptive finite element methods lend themselves to parallel computation. New results on parallelization of adaptive methods suggest that significant speed-up times in large scale computation can be realized with parallel-adaptive algorithms. This is a relatively new field, but its potential in improving speed and reliability of engineering computations is very great.

6. Adaptive *p* and *h-p* methods are now under study which exhibit exponential rates of convergence. Work needs to be done to control and minimize the computational overhead using such methods. The possibility of delivering exponential convergence rates in practical engineering simulation is very significant; if achieved, and results presented at the Workshop clearly show that such convergence rates are not uncommon in *p*- and *h-p* methods, then these approaches may emerge as the most important modeling methods available. In principle, such methods can give results of a specified accuracy on a machine with a fixed memory, that cannot be attained by any conventional (low order, unadapted) method.

7. Mesh generation techniques continue to be a crucial issue in computational modeling. Much of current mesh technology was designed for structured meshes and conventional finite difference and finite element methods. New developments in adaptivity make use of these techniques obsolete and a new era of mesh generation is imminent. New techniques for efficient generation and adaptation of two and three dimensional meshes which are based on developing ideas in adaptivity, *p*- and *h-p* representation, moving mesh methods embedding techniques, zoning methods, and general mesh optimization concepts are emerging in contemporary

computational mechanics.

The Workshop established that reliability of computational modeling is an issue of growing concern across all fields of engineering and stands as a key issue to future progress in computational mechanics. Fortunately, many new developments in numerical modeling and analysis are emerging that should have a significant impact on reliability in computational mechanics. Future research directions should focus on work in making more precise and rigorous guidelines for the mathematical modeling process itself, on *a priori* and on *a posteriori* error estimation, on developing adaptive methodologies, on parallel computing, and on mesh generation. On a more applied level, the sensitivity of users of computational methods to issues of reliability must be sharpened if the full value of computer modeling is to be realized, and this can be done if educators, developers, and users of computational modeling insist on a higher level of reliability and on a means to assess it.

J. Tinsley Oden
The University of Texas at Austin

STATEMENT "A" per Dr. R. Lau
ONR/Code 1111MA
TELECON 5/16/90 VG

This workshop was supported in part by Grant #N00014-89-J-1772, issued by the Office of Naval Research.

ADAPTIVE GRIDS FOR COUPLED VISCOUS FLOW AND TRANSPORT

K.C. Wang and G.F. Carey

University of Texas at Austin

Abstract

An adaptive grid refinement and coarsening scheme in two dimensions for biquadratic elements is developed. This has been incorporated in a two-dimensional finite element program for steady and transient generalized Newtonian flow problems. Numerical results are given for Navier-Stokes and power law non-Newtonian flow calculations as well as coupled fluid flow and transport problems including free surfaces, theromocapillary flows, die swell, and electro-rheological flows.

1 INTRODUCTION

The need to obtain accurate finite element solutions efficiently for viscous flow calculations, has stimulated the development of adaptive refinement procedures for automatically improving the grid and solution. Ideally, the final mesh so obtained should be graded into regions where the solution and its derivatives vary significantly so that the error on the domain is uniformly small. Research on this subject primarily concerns the following topics: Development of reliable *a posteriori* error estimates to determine where refinement is warranted (e.g. see Babuška and Rheinboldt [1]); development of refinement data structures for efficient storage (e.g. see Bank and Sherman [5]); and incorporation of fast iterative solution procedures (e.g. see McCormick [22], Carey and Humphrey [8]). Other contributions in this area and related references are collected in the monographs edited by Shephard and Gallagher [31], Ghia and Ghia [17], Lohner *et al.* [20], Babuška *et al.* [3], Develoo *et al.* [14], and Carey *et al.* [12].

In the present study we focus on the development of a refinement data structure and apply this to viscous flow and transport problems including applications with free surfaces. Both Navier-Stokes problems and power law non-Newtonian flows are considered. Here we present a refinement data structure for 9-node biquadratic elements. Previous studies (e.g., Babuška and Rheinboldt [2], Rheinboldt and Mesztenyi [28], Bank and Sherman [5], Sharma and Carey [30], Jiang and Carey [19], Ludwig *et al.* [21]) have considered the implementation of adaptive refinement algorithms for triangular and bilinear elements. In a different setting, there has been work on solution enhancement by increasing polynomial degree (p-refinement) as in the studies of Szabo [33]. One can also combine mesh refinement with polynomial enhancement (Rank [27]). However, the logic and use of these hierarchic bases and the approach is substantially different from the present work. The data structure employed is an extension of that developed by Bank and Sherman [5] for linear triangles and more recently by Sharma and Carey [30] for bilinear quadrilaterals. This data structure stores a small amount of data for efficient construction of the system and solution. For steady state flow problems, nonlinear solution is carried out by Newton iteration in conjuction with a line-search strategy. For flows at higher Reynolds numbers, incremental continuation in the Reynolds number is also employed. The procedure has also been extended for analysis of transient flows so that both refinement and coarsening of the grid is permitted during the time-stepping procedure. In this way fine features of the flow structure can be followed throughout the flow field.

2 Viscous Flow and Transport Equations

The class of problems considered in this work are those described by the viscous flow and energy transport equations. The fluid motion is further assumed to be two-dimensional (or axisymmetric), and laminar. The fluids are considered to be incompressible, Newtonian, or non-Newtonian (such as power-law or Bingham fluids). The governing equations are derived from the basic physical principles of conservation of mass, linear momentum and energy, together with constitutive equations which relate the stress to the rate of deformation, heat flux to the temperature, and density to the temperature. The resulting equations for steady flow can be written in Cartesian tensor form as follows:

$$\rho u_j u_{i,j} = \tau_{ij,j} + \rho g_i \tag{1}$$

$$u_{i,i} = 0 \tag{2}$$

$$\rho c_p u_i T_{,i} = -q_{,i} + \Phi + \rho q_s \tag{3}$$

where ρ is density, u_i is velocity, g_i is the gravity vector, τ_{ij} is the total stress, c_p is the heat capacity, T is temperature, q is the heat flux, Φ represents viscous dissipation, and q_s is the volumetric heat source. In this study we consider flow conditions where Φ is negligible.

The above equations represent, in general, a coupled system that requires boundary conditions on both the fluid motion and the energy transport. The necessary boundary conditions are of the standard type and consist of specified velocities or tractions for the momentum equations and specified temperature or heat flux for the energy equation. There is one other type of boundary condition requiring further discussion here which concerns the condition along a free surface. The most common example of this type of problem is exit flow of a jet as in the die-swell example considered later. Along the free surface the normal component of the surface traction balances the external pressure and surface tension and the tangential component is zero. That is,

$$\tau_{ij} n_j = p_a n_i + \gamma \left(\frac{1}{R_1} + \frac{1}{R_2} \right) n_i \tag{4}$$

$$\tau_{ij} s_j = 0 \tag{5}$$

where n_j and s_j are unit vectors normal and tangential to the free surface, p_a is the ambient pressure, γ is the surface tension, and R_1, R_2 are the principal radii of curvature. In this work, surface tension effects are considered to be negligible. Also, without loss of generality, the ambient pressure may be set to zero. Thus, the appropriate free surface boundary conditions correspond to a vanishing of the normal and tangential stresses along the free surface. However, the shape of the free surface is not known *a priori* and must be located as part of the solution.

Part of the present research is directed towards the analysis of shear thinning fluids and the related solution algorithms in finite element models. The term "generalized" Newtonian fluid has also been used for this type of non-Newtonian fluid in the literature by Bird et al. [6], Gartling [15] and others. The form of the constitutive equation for this type of non-Newtonian fluid is given by

$$\tau_{ij} = -p\delta_{ij} + 2\eta D_{ij} \tag{6}$$

where τ_{ij} is the total stress tensor, p is the pressure, δ_{ij} is the unit tensor, D_{ij} is the rate of deformation tensor

$$D_{ij} = \tfrac{1}{2}(u_{i,j} + u_{j,i}) \tag{7}$$

and the apparent viscosity η is a function of the shear rate. A variety of models for η have been proposed and correlated with experimental data. In this work we consider power-law models since they have relatively simple form but are extensively used in industry. In this case

$$\eta = K I_2^{(n-1)/2} \tag{8}$$

where K is the consistency factor, $I_2 = \frac{1}{2}tr(D^2)$ is the second invariant of D_{ij}, and $n > 0$ is the power law index. One of the significant features of power law fluids is the shear thinning effect, in which the apparent viscosity decreases with increasing shear rate, when $0 < n \leq 1$.

For an isotropic material, the heat flux can be written as $q_i = -kT_{,i}$ where $k > 0$ is the thermal conductivity. For non-isothermal flows, an extended form of the Boussinesq approximation (Gray and Giorgini [18], McLay [23]) is used to accommodate buoyancy forces. This allows the fluid properties to be functions of the thermodynamic state and density to vary with temperature according to $\rho = \rho_o[1 - \beta(T - T_o)]$ where subscript o refers to a reference state, and $\beta > 0$ denotes the volume expansion coefficient. For isothermal flows $\rho = \rho_o =$ constant which is simply the assumption of incompressibility.

3 Finite Element Formulation

Consider a flow domain Ω where $\partial\Omega$ is the total boundary enclosing the domain, $\partial\Omega_u$ and $\partial\Omega_T$ are parts of the boundary with specified velocity components and temperature respectively. On $\partial\Omega - \partial\Omega_u$ and $\partial\Omega - \partial\Omega_T$ traction and flux or mixed boundary conditions apply. Beginning with the basic equations (1)–(3) and constitutive equations (6)–(8), a Galerkin-based weighted residual method, (e.g., see Carey and Oden [9]) leads to a weak form of the basic equations

$$\begin{aligned}\int_\Omega (\rho_o u_j u_{i,j} v_i + \tau_{ij} v_{i,j} - p v_{i,i} + \rho_o g_i T v_i) d\Omega &= \int_\Omega \rho_o g_i (1 + \beta T_o) v_i d\Omega \\ &+ \int_{\partial\Omega - \partial\Omega_u} \tau_{ij} v_i n_j ds \qquad (9)\end{aligned}$$

$$\int_\Omega q u_{i,i} d\Omega = 0 \qquad (10)$$

$$\begin{aligned}\int_\Omega (\rho_o c_p u_i T_{,i} w + k T_{,i} w_{,i}) d\Omega &= \int_{\partial\Omega - \partial\Omega_T} k T_{,i} n_i w ds \\ &+ \int_\Omega \rho_0 q_s w d\Omega \qquad (11)\end{aligned}$$

for all admissible test functions v_i, q, w with $v_i = 0$, $w = 0$ on those parts of the boundaries $\partial\Omega_u$ and $\partial\Omega_T$ where u_i and T are specified. The surface integrals involve the applied surface stresses (tractions) and heat flux and permit these to be introduced as natural boundary conditions for the variational problem.

In a mixed finite element formulation we approximate the velocity, pressure and temperature using piecewise-polynomials Let V^h, P^h and θ^h be the approximation spaces so determined. The finite element problem is to find $u_h \epsilon V^h$, $p_h \epsilon P^h$, and $T_h \epsilon \theta^h$ satisfying the essential boundary conditions and such that

$$\begin{aligned}\int_{\Omega_h} [\rho_0 (u_j)_h (u_{i,j})_h (v_i)_h + (\tau_{ij})_h (v_{i,j})_h] d\Omega &- \int_{\Omega_h} [p_h (v_{i,i})_h - \rho_0 g_i T_h (v_i)_h] d\Omega \\ &= \int_{\partial\Omega_h - \partial(\Omega_u)_h} (\tau_{ij})_h (n_j)_h (v_i)_h ds \\ &+ \int_{\Omega_h} \rho_0 g_i (1 + \beta T_o)(v_i)_h d\Omega \qquad (12)\end{aligned}$$

$$\int_{\Omega_h} q_h (v_{i,i})_h d\Omega = 0 \qquad (13)$$

$$\int_{\Omega_h} [\rho_0 c_p (u_j)_h (T_{,i})_h w_h + (kT_{,i})_h (w_{,i})_h]\, d\Omega = \int_{\partial\Omega_h - (\partial\Omega_T)_h} (kT_{,i})_h (n_i)_h w_h dS + \int_{\Omega_h} \rho_0 q_s w_h d\Omega \qquad (14)$$

hold for all admissibl· ·unctions $v_h \epsilon V^h$, $q_h \epsilon P^h$ and $w_h \epsilon \theta^h$ and with τ_{ij} given by (6) – (8). Introducing finite element expansions for u_h, p_h and T_h and finite element test bases for v_h, q_h and w_h into (12)–(14) and integrating, we obtain a system of non-linear algebraic equations which can be written in matrix notation as

$$\begin{cases} C(u)u + Ku + BT - Qp = F \\ -Q^T u = 0 \\ D(u)T + LT = G \end{cases} \qquad (15)$$

The nonlinear system above may be solved in either fully coupled or iteratively decoupled form. For the coupled flow and transport considered here we use the fully coupled form since this is applicable to a wider class of flows. The nonlinear algebraic system (15) can be expressed conveniently as

$$g(z) = 0 \qquad (16)$$

where z represents the vector of nodal velocities, pressures and temperatures. In the present analysis, the biquadratic nine-node velocity, bilinear four-node pressure element is employed. (This element does not admit oscillatory spurious pressure modes.) For flow at low and moderate Reynolds numbers the nonlinear system can be solved efficiently by Newton iteration of the associated Jacobian system

$$J^i \delta^i = -g^i \qquad (17)$$

where δ^i is the correction in the solution vector and the Jacobian matrix J is evaluated at the solution for the current iterate i. This solution procedure can be easily combined with incremental continuation in the Reynolds number or arc-length continuation for more demanding nonlinear flows.

In the case of the transient viscous flow problem, the nonlinear algebraic system (16) is replaced by the semidiscrete system

$$M\dot{z} + g(z) = 0 \qquad (18)$$

which is integrated numerically in time from specified initial data.

In the following studies we also consider power law fluids, which again lead to steady and transient approximate problems of the form given in (17) – (18), where now the nature of the nonlinearity also depends upon the index for the fluid. For low power law index the fluid exhibits shear thinning and this can cause problems with the convergence of the Newton iteration. If a line search strategy is included in the algorithm, then convergence with the shear thinning fluids can be achieved.

Solution of the linear Jacobian system in (17) or the corresponding linear implicit systems arising from the semidiscrete formulation (18) is achieved using a frontal solver. An integral part of any frontal solver is the prefront algorithm which establishes pointers related to the element connectivity to be used in the frontal solution. In the present case the mesh is unstructured (containing constrained nodes) and more importantly, the mesh changes as the refinement procedure is carried out. In the case of the steady flow calculations, the initial mesh is coarse and mesh refinement is monotonic from this initial coarse grid. This means that we consider only refinement of the grid when the steady state computations are required, whereas both refinement and recombination are involved in the unsteady computations. Following each refinement step the prefront scheme is again called to generate the new destination vectors for the sparse solution algorithm. In this way the efficiency of the frontal scheme is not degraded by the adaptive refinement.

4 Adaptive Mesh Refinement

Of paramount importance to an effective data structure supporting an adaptive mesh refinement scheme are: 1) storage requirements; 2) CPU time; and 3) ease of implementation. The data structure described here stores a small amount of data for efficient construction of this system and numerical solution, thereby providing a good balance between storage and computation overhead. Moreover, the data structure is logically separate from the finite element analysis and hence can be incorporated into other existing finite element codes.

The basic data structure is a quadtree, defined by refining quadrilateral elements to a quartet of subelements in any given refinement step. These subelements may be later individually subdivided to further quartets. A pointer system links different levels in the quadtree so determined (Carey, Sharma, and Wang, [11]). To ensure a smooth transition from coarse to fine grid size in the mesh and to simplify the data structure, no two neighboring elements are permitted to have a level difference greater than one. This strategy is commonly employed in other adaptive refinement codes using linear elements for similar reasons. The adaptive refinement procedure, therefore, is recursive: If the error indicator for a given element is large and the element is to be refined, then neighbor element information is needed to determine the level of the adjacent elements. When the need to refine implies that the level rule would be violated, the corresponding neighbor element must first be refined. This in turn implies that its neighbors must be interrogated and so on recursively until no level violation is encountered.

Error indicators based on the solution for the current mesh guide the refinement procedure. Traditional error estimates in finite element analysis consist of *a priori* global bounds which yield theoretical asymptotic rates of convergence depending on the characteristic mesh size. For adaptive refinement, local computable error estimators and indicators are needed (for example, see Babuška and Rheinboldt, [2], Bank [4]). Several forms of error indicators are currently in use. The most popular strategies employ element interpolation error estimates or residuals. The residual associated with the weak statement of a given problem represents the amount by which the differential equation is not satisfied locally (in a sense similar to truncation error in finite difference approximation). Since biquadratic elements are used in the present study, the L_2 norm of the interior element residual is easy to calculate and has been used in the numerical experiments described later. (For potential problems in which linear elements are used it is necessary to compute the interface jumps, e.g., see Sharma and Carey [30].) In this study we use Gaussian quadrature to compute the L_2 norm of the residual for each element; this is then normalized according to element size. The mean and standard deviation of residual norms are then computed. If the normalized error norm of an element exceeds the mean error by an amount more than the standard deviation, we refine the element. For problems involving energy transport, the residual norms of the momentum and the energy equations are computed separately. The mean and standard deviation of residual norms for the momentum and the energy equation are also computed separately. If the normalized error measure for either the momentum or the energy equation of an element exceeds the corresponding mean error by an amount larger than the corresponding standard deviation, we refine the element. The procedure stops when the mean residual is reduced to the level of the best residual in the initial mesh, or a specified fraction of its original value. The solution changes may also be monitored as part of the stopping criterion.

One of the most critical points in the design of adaptive refinement software is the data structure. A poorly designed data structure may necessitate excessive storage and not lend itself to efficient refinement. In recent years several data structures have been developed and implemented into computer packages. For example, the data structure proposed by Rheinboldt and Mesztenyi [28] employs a general labelled tree and poses no restrictions on the order of irregularity (no level restriction between neighboring elements) of the mesh. It has a set of specialized algorithms which allow efficient traversal of the tree structure and linear solution based upon nested dissection techniques (see George [16]). The data structure presented by Bank and Sherman [5] provides efficient storage and permits adaptive refinement of linear triangles together with a multigrid solution algorithm. Rivara [29] uses a "molecular List" structure for adaptive conforming triangulation with a multigrid solution method. The data structure presented by Diaz et al. [13] is a modification of that by Rheinboldt and Mesztenyi, to support both mesh refinement and mesh recombination for time dependent problems.

The principal information describing our quad-tree data structure for biquadratics is contained in two integer arrays of dimension $8 \times \max e$ and $2 \times \max n$, respectively, where $\max e$ and $\max n$ are the maximum number of elements and nodes respectively in the mesh. The columns of the first array contain pointers and nodal numbers in order to track connectivity of new elements created by refinement. Consider an element with nodal numbers as shown on the left of Figure 1. After refinement the nodal numbers of the refined quartet are shown on the right part of the figure. Columns j through $j+3$ of the first array contain the information for the quartet, arranged as shown in Figure 2. Here F, L, and MF denote the father, level and macro father (user-defined element) of the quartet; S_j is the son pointer of element j; indices 1-9 are nodal numbers of the father element, E_k, $E_k\,'$ define mid-side nodes along edge k of elements j and $j+1$ where $[k = \bmod\ (j-1,\ 4)+1]$; N_k is the normal node between j and $j+1$ (with $k = 1, 2, 3$ or 4) and C_k is the center node of element j [with $k = \bmod\ (j-1, 4)+1]$. The second array contains information related to the nodes. Using this Table the type of node (boundary, constrained, etc.) and neighbor information can easily be determined. (See Carey *et al.* [11] for a more detailed description of these two arrays.)

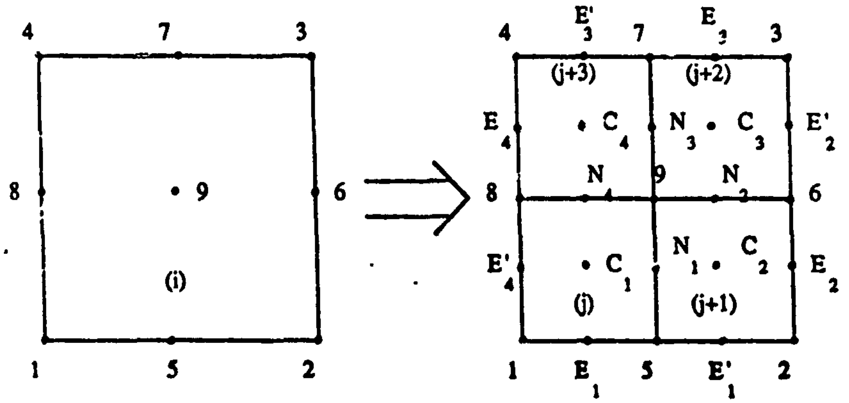


Figure 1: Biquadratic element numbering and node types for refinement scheme (see also Figures 2 and 3)

The actual refinement process requires specification of the entries of these two arrays. Some of the data for these arrays is standard and trivially known from the parent element data. The remaining information can be computed directly from the existing data structure. This standard information includes: nodal numbers of a given element, neighbors of an element, father of an element, level of an element, and node fathers of a given node.

As mentioned earlier, this refinement procedure can be integrated into an existing finite element code without difficulty. In this work, we interface this refinement procedure with a 2-D incompressible generalized Navier-Stokes finite element code, which uses a frontal solver technique to solve the Jacobian system for Newton iteration. Following are modifications we have made in order to interface with this refinement procedure.

(1) Before the refinement procedure can be invoked, we need to initialize the macro-edge

column j	column j+1	column j+2	column j+3
F	E_1	E'_1	N_1
L	E_2	E'_2	N_2
MF	E_3	E'_3	N_3
C_1	E_4	E'_4	N_4
C_2	4	7	3
C_3	8	9	6
C_4	1	5	2
S_j	S_{j+1}	S_{j+2}	S_{j+3}

Figure 2: Connectivity array for quad-tree refinement data structure.

array which defines the neighbors of macro elements.

(2) An array to store son pointers of macro elements is also needed.

(3) Following each refinement step, the pre-front scheme is again called to generate the new destination vectors for the frontal sparse solution algorithm. The pre-front routine utilizes "nicknames" for the nodes consisting of the node number and degrees of freedom of the node. Note that there is no degree of freedom associated with exposed nodes and we are using pre- and post-matrix multiplications to handle the constraint condition on exposed nodes. Therefore, the nickname of an exposed node should be replaced by the nickname of a node of the father element used to interpolate the value at this exposed node. Consider a finite element grid as shown in Figure 4. Nodes 22 and 23 are exposed nodes following a refinement. Solutions at nodes 22 and 23 are obtained through interpolation of nodes 2, 3, and 6. Therefore, we use the nickname for node 3

USER	EXPOSED	NORMAL	BOUNDARY	CENTER
0	-IVF1	IVF1	IVF1	IVF1
ibc(i)	F	IVF2	ibc(i)	0

Figure 3: Vertex array for quad-tree refinement data structure

in place of the nickname for node 23 and nickname for node 2 in place of nickname for node 22.

(4) For boundary nodes with Dirichlet boundary condition specified, interpolation is used to determine the boundary data of nodes created via refinement.

(5) During the matrix assembly phase a matrix pre- and post- multiplication procedure has been implemented to handle exposed nodes. This procedure can best be explained by means of an example. Again consider the finite element grid as shown in Figure 4. Nodes 22 and 23 are exposed nodes following refinement to the quartet on the right. For the biquadratic basis considered here, the constraint equation for node 23 of element 3 is simply $\boldsymbol{u}_{23} = 3/8\boldsymbol{u}_2 + 3/4\boldsymbol{u}_6 - 1/8\boldsymbol{u}_3$. Using the approach in Carey [7], we introduce an elementary transformation matrix $\boldsymbol{E}$ relating constrained and unconstrained nodal values of an element by $\boldsymbol{u}_c = \boldsymbol{E}\boldsymbol{u}$ with $E_{ij} = \delta_{ij}$, if i is a regular node and if i is the constrained node above, $E_{i,i-4} = 3/8$, $E_{i,i-3} = 3/4$, $E_{i,i} = -1/8$ with the rest of the entries in $\boldsymbol{E}$ zero. Formally the constrained element contributions for $\boldsymbol{K}$ and $\boldsymbol{F}$ can be computed as $\boldsymbol{K} = \boldsymbol{E}^T\hat{\boldsymbol{K}}\boldsymbol{E}$, $\boldsymbol{F} = \boldsymbol{E}^T\hat{\boldsymbol{F}}$ where $\hat{\boldsymbol{K}}$, $\hat{\boldsymbol{F}}$ are the usual contributions for the unconstrained biquadratic. Of course, in practice this matrix multiplication is not

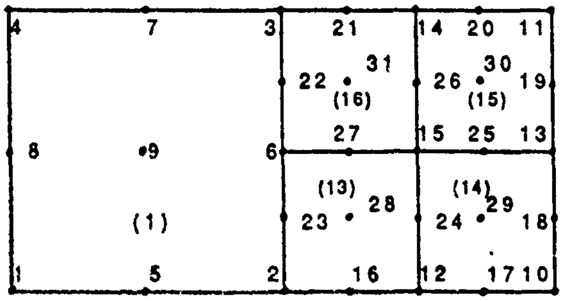


Figure 4: Interface between refined and unrefined elements containing "exposed" nodes 22 and 23 to be constrained.

carried out but instead the element contributions $\boldsymbol{K}$, $\boldsymbol{F}$ are constructed directly.

(6) Using the solution on the previous mesh with interpolation at the new grid points yields an initial vector for the next iteration on the new mesh. Alternatively, a local projection can be introduced to improve the starting iterate (Carey and Seager [10]).

(7) During pre-front and matrix assembly phases, instead of processing elements sequentially, we follow the pointers to process other children of the quartet first before processing the next element in sequential order. Again, let examine Figure 4. Element 2 has been refined to produce new elements 13 through 16. Now element 2 is inactive and after processing element 1, we process elements 13 through 16 before processing element 3. In this fashion, the front width is perturbed only by a small increment from the original mesh front width.

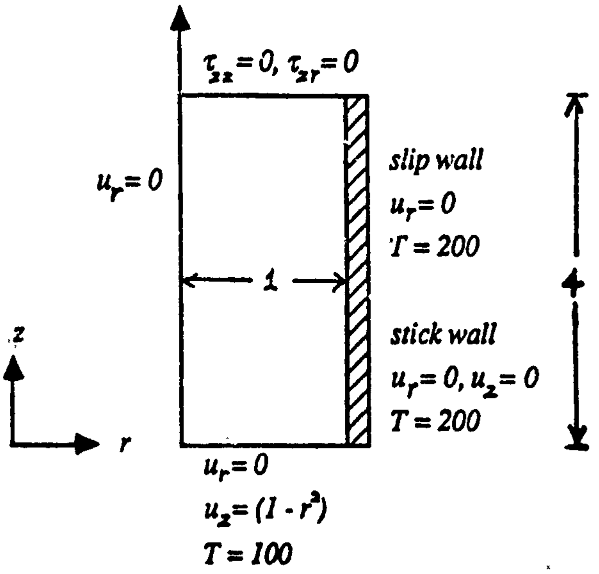


Figure 5: Geometry of stick-slip problem with hot wall.

5 Numerical Results

5.1 Stick Slip Problem

We present here some representative results for coupled viscous flow and heat transfer problems. The first test problem is a stick-slip problem for axisymmetric viscous flow in a pipe with no-slip condition on the stick wall (Figure 5). Fluid enters the cylinder as plug flow with temperature T_0 and encounters the wall at temperature T_1. Thus there are singularities in both the temperature field at the entry wall and in the flow field at the stick-slip point. Since these singularities have a pronounced effect on the local fluid flow and heat transfer fields, they provide a suitable mechanism for testing the adaptive refinement algorithm.

Beginning from a uniform coarse grid, the mesh is locally refined through several levels to produce the grids shown in Figure 6. We see that the grid is refined locally into the two singular regions with more pronounced refinement occurring at the stick-slip interface. In these refinement calculations, the refinement test was based on residual error indicators computed from both the viscous flow and energy transport equations.

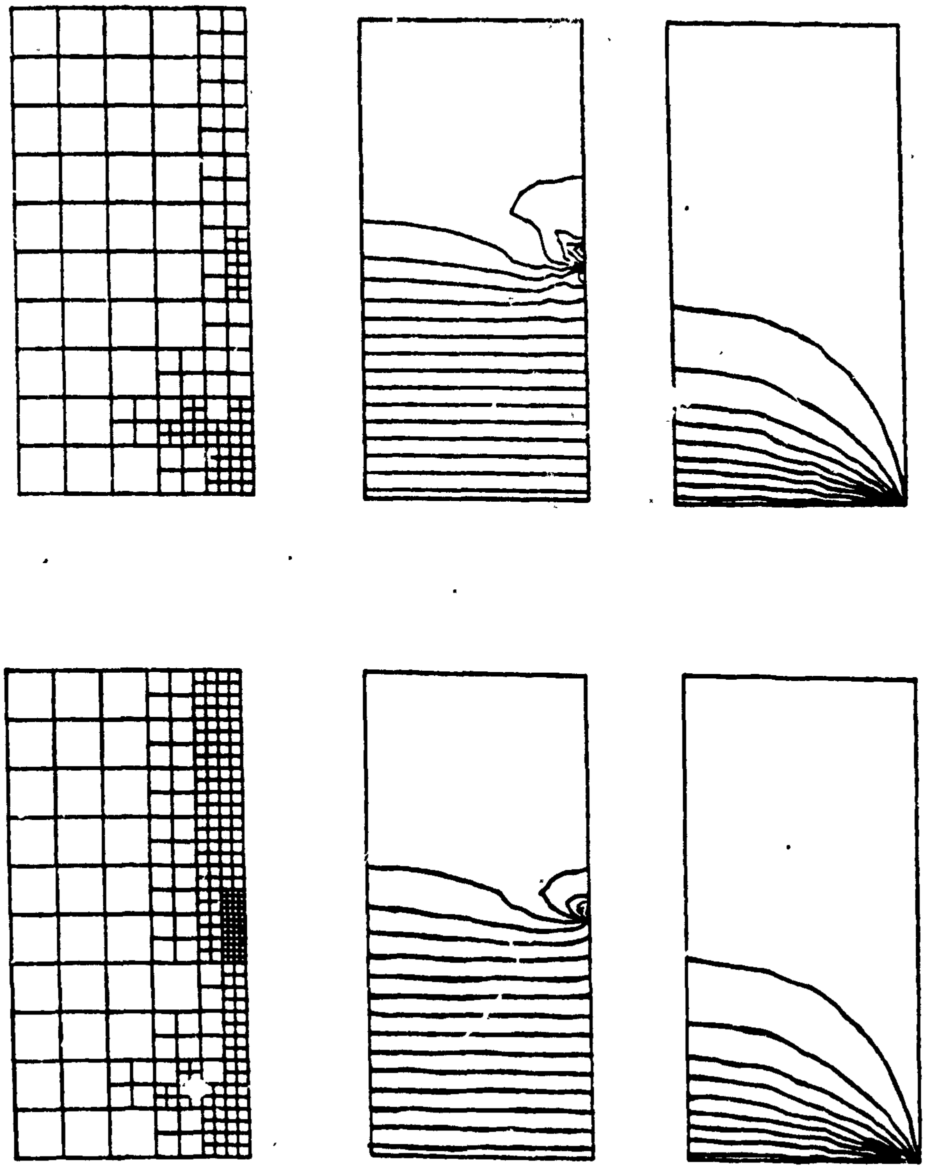


Figure 6: Finite Element mesh, pressure and temperature contours after 2 and 3 refinements.

$n = 1.0$			
	graded mesh	2 ref.	3 ref.
error estimate	-	.2382	.3641
refinements	-	.2108	.4198
total time	42.056	22.310	35.376
r_f/r_i	1.121	1.130	1.128

$n = 0.8$			
	graded mesh	2 ref.	3 ref.
error estimate	-	.2368	.3594
refinements	-	.1887	.3724
total time	39.417	21.382	28.584
r_f/r_i	1.092	1.101	1.099

$n = 0.6$			
	graded mesh	2 ref.	3 ref.
error estimate	-	.2362	.3662
refinements	-	.1886	.4042
total time	38.099	19.914	23.303
r_f/r_i	1.060	1.067	1.066

Table 1: Break down of cpu time and die-swell ratio for Newtonian jet ($n = 1.0$) and power-law jets ($n = 0.8$ and $n = 0.6$).

5.2 Die Swell Problem

Similar calculations were performed for the die swell problem with a power law fluid. In this case, in addition to the singularity at the tip of the stick wall, there is a free surface which must be determined iteratively as part of the solution algorithm during refinement. Moreover, as refinement takes place at elements adjacent to the free surface, new nodes are introduced on the free surface, and this has to be treated appropriately in the algorithm. The free surface configuration and adapted mesh are show in Figure 7 following three levels of refinement from an initial uniform grid of 50 elements in a rectangular domain. The final adaptively-computed swell ratio for a Newtonian fluid and two power-law fluids as compared to a graded mesh with 112 elements is shown in Table 1.

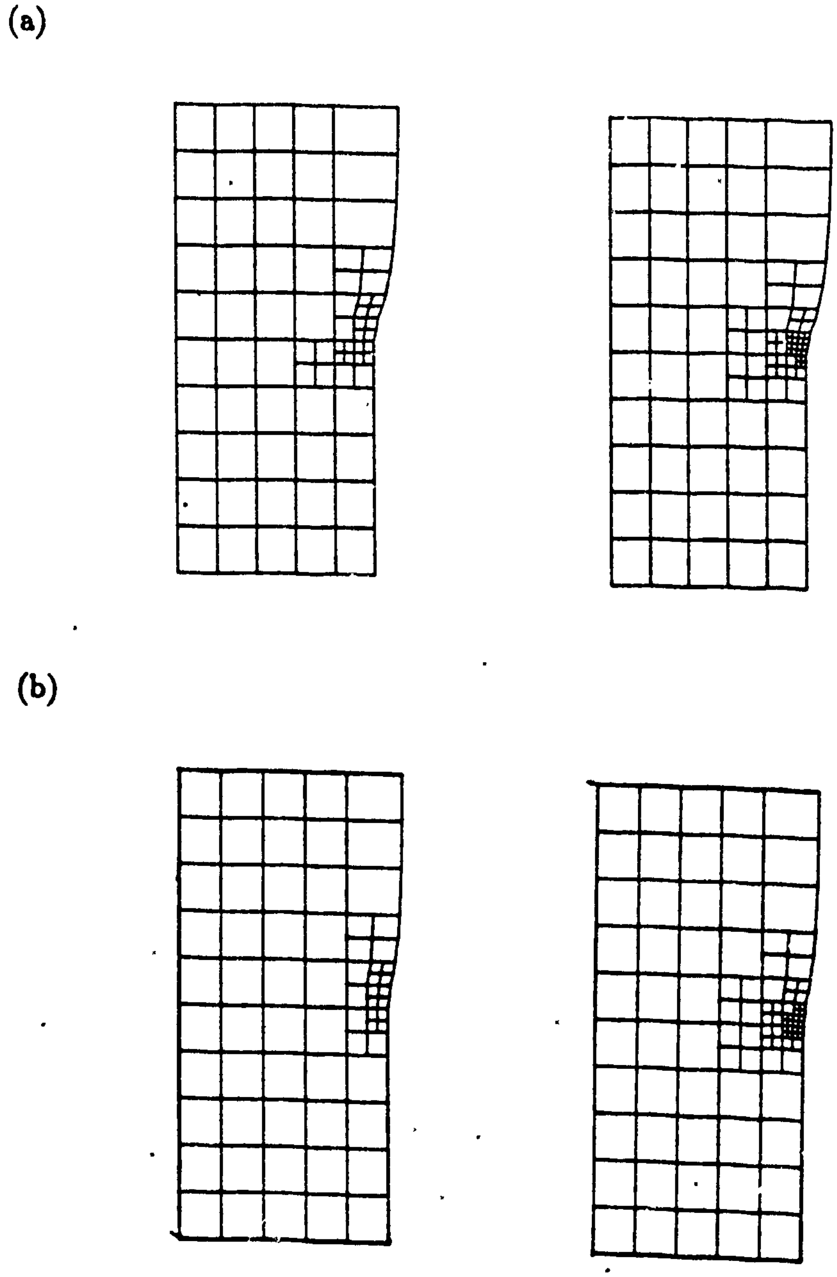


Figure 7: Final mesh and die swell, (a) $n = 1.0$, (b) $n = 0.6$.

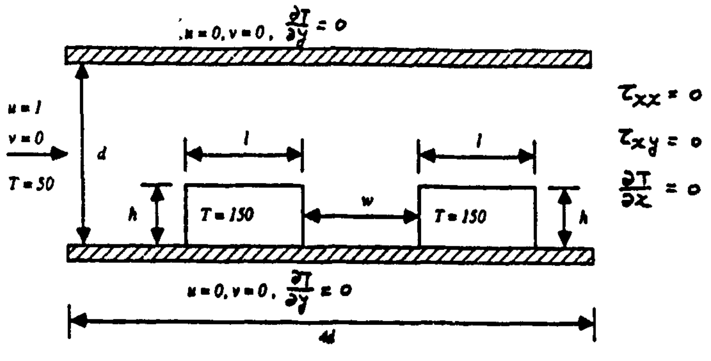


Figure 8: Tandem of heated rectangular blocks.

5.3 Electronic Cooling

The next example is motivated by electronic cooling applications and considers the convective heat transfer of a tandem of heated blocks mounted on the wall of an adiabatic channel (Figure 8). First the flow field of a single block configuration is examined. Figure 9 shows the velocity field for a uniform mesh and following adaptive refinement at Reynolds number equal to 200. The effect of adaptive refinement can be seen clearly in the suppression of local oscillations in the flow. The geometry and initial finite element mesh for a two-block configuration are shown in Figure 10. Figures 11 and 12 give the temperature profiles for block spacing $w = l$ and $w = 2l$ using adaptive refinement starting from the initial uniform mesh shown in Figure 10. The effect of choice of fluid, "sheltering" of the second block by the first and the influence of spacing can be investigated for such applications as microelectronic cooling. Finally, the approach has been extended to transient problems where, in addition to mesh refinement, mesh recombination is carried out. For example, if a sharp solution gradient is convected across the domain, then elements in front of the advancing layer will be refined, and elements behind the layer will be recombined as the solution is integrated forward in time.

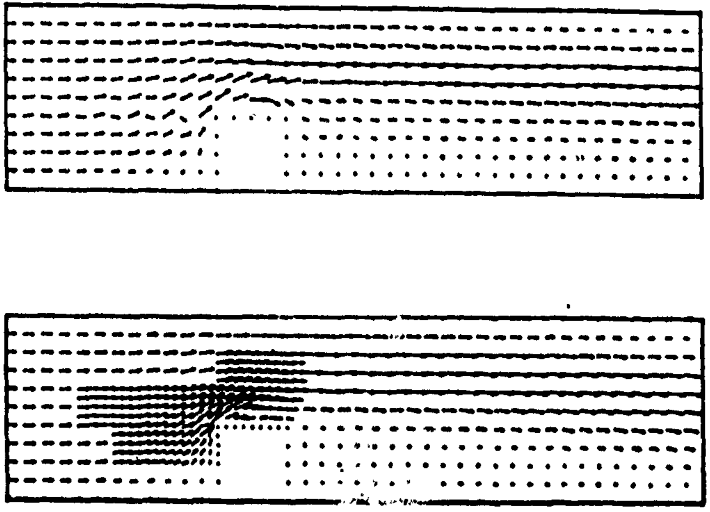


Figure 9: Velocity vectors for initial and refined grid (Re=200).

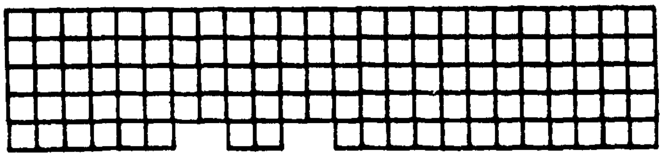


Figure 10: Initial grid for double block configuration.

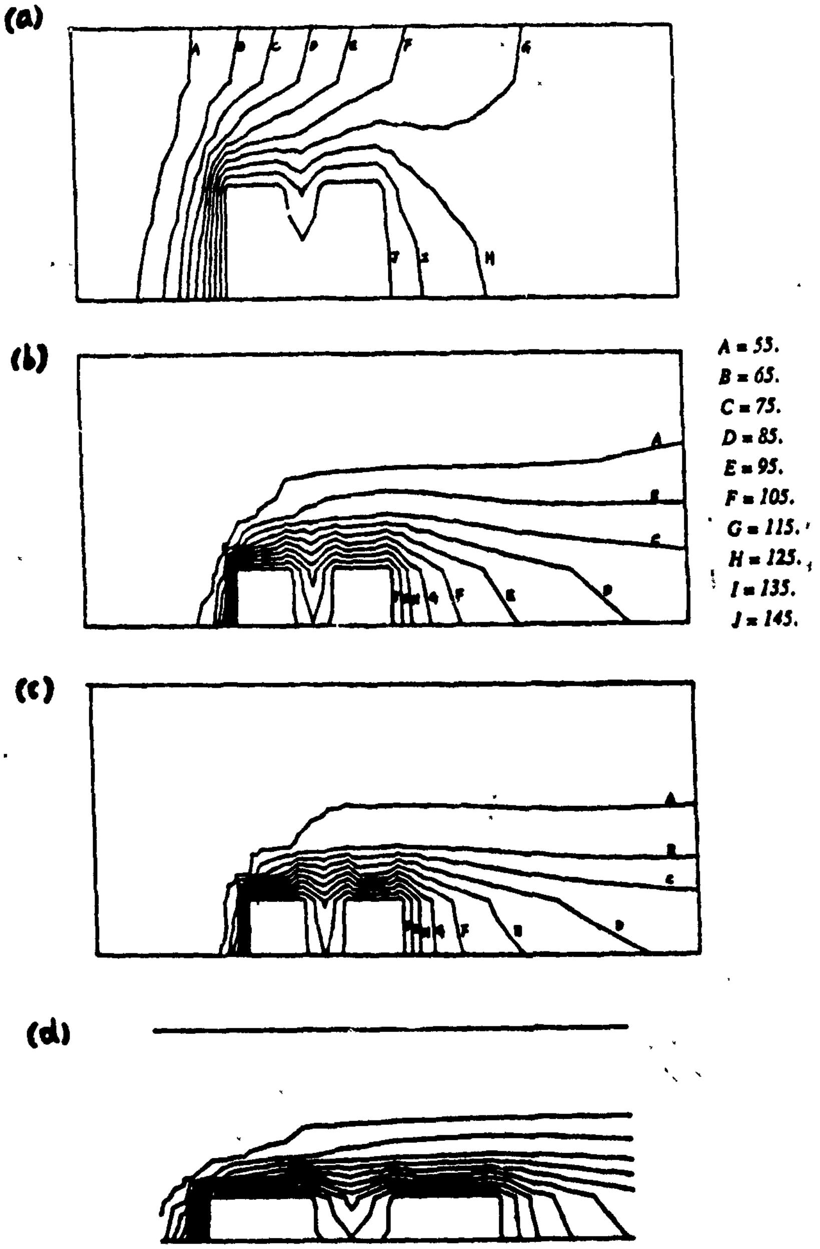


Figure 11: Temperature contours for $w = \ell$, (a) $Re = 10$, (b) $Re = 100$, (c) $Re = 200$, (d) scaled plot ($Re = 100$).

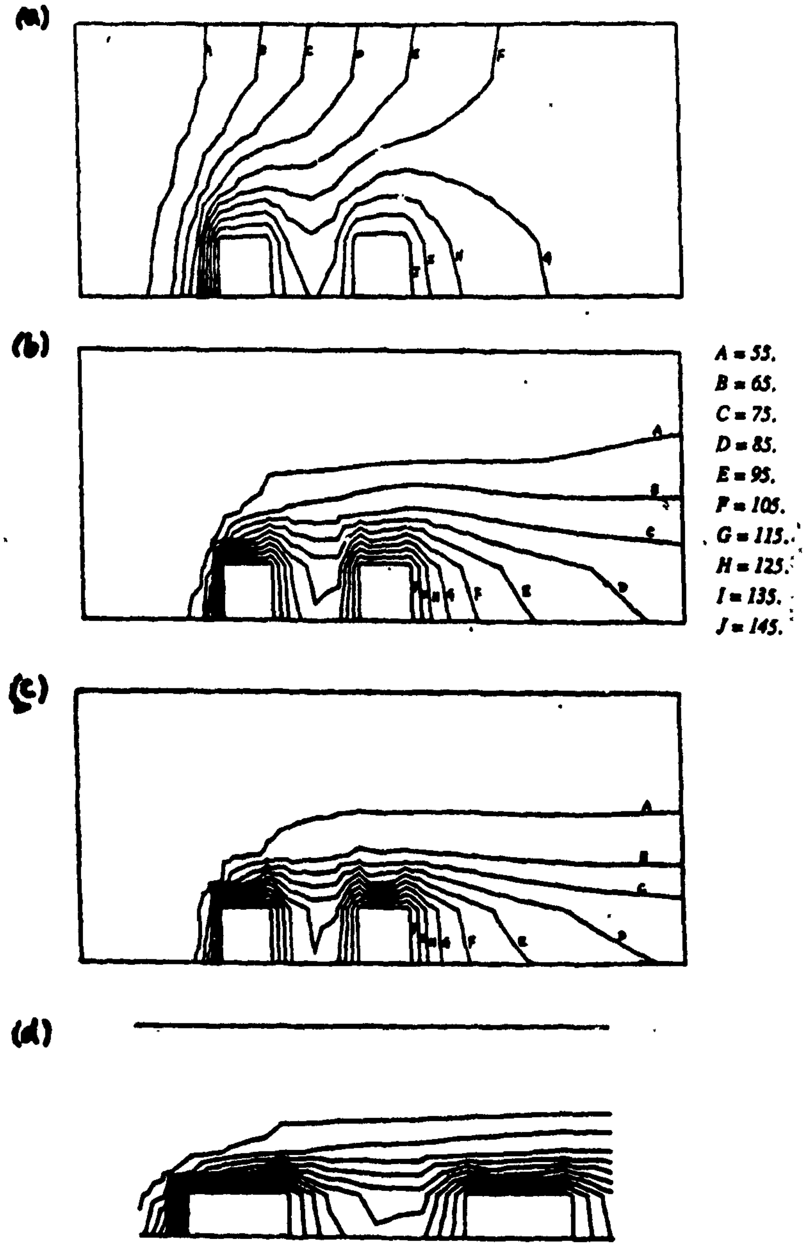


Figure 12: Temperature contours for $w = 2\ell$, (a) $Re = 10$, (b) $Re = 100$, (c) $Re = 200$, (d) scaled plot ($Re = 100$).

5.4 Thermocapillary Flow

The variation of surface tension with temperature can be a dominant feature in certain coupled fluid flow and heat transfer problems. This thermocapillary effect is important in welding (e.g., see Ostrach [25], Mclay and Carey [24]). In these high Marangoni number flows ($Ma = RePr$) surface tension drives the flow at relatively high Reynolds numbers (e.g., $Re = 0(10^3)$) in the welding applications. Previous numerical studies of thermocapillary flows for a cavity have given reliable results at low Re and Ma (Strani *et. al.* [32]). However, at higher Ma, the solution is very sensitive to the choice of mesh. This class of problems, therefore, provides a good test of the ability of adaptive procedures to circumvent these grid-related issues.

The test problem is a unit square cavity open at the top and containing an incompressible, viscous fluid. The vertical walls are heated at respective temperature T_L and T_R and the bottom is adiabatic. At the free surface the normal heat flux is zero. The differential heating of the side boundaries produces a thermal gradient. Since the temperature of the fluid at the free surface varies, the thermocapillary effect enters with

$$\frac{\partial u}{\partial y}(x,1) = -\frac{\partial T}{\partial x}(x,1)$$

giving the surface tension boundary condition for the momentum equations. Under the action of the surface tension, a convective flow pattern develops to a steady state solution. The corners $(0,1)$ and $(1,1)$ are singular points in the flow (as is also the case in the familiar driven cavity problem). At high Ma and Re, the solution has strong layers adjacent to the top surface and particularly near the cold corner (here taken to be $(1,1)$). We computed the solution on a coarse uniform grid but the layer structure was not captured.

This problem is also computed by Zebib *et. al.* [35] using a finite difference scheme based on the methods in Patankar [26]. Mesh refinement studies on uniform grids of size 65 × 65 and 80 × 80 indicate the presence of the layers.

In the present calculations, we begin from a uniform rectangular grid of size (20 × 10) and the scheme adaptively refines the grid during solution. The error indicators correctly locate the corner and layer regions and refine accordingly. As an example, the mesh after

3 refinement steps is given in Figure 13 for $Re = 10000$, $Pr = 0.1$. The corresponding surface velocity profiles on adaptive meshes for $Pr = 0.1$ at $Re = 1000$, 5000 and 10000 are shown in Figure 14 and the surface temperature profiles in Figure 15. As the solution and mesh are developed iteratively from coarser to finer grids, the nonlinear solution scheme is both efficient and robust. The calculation was repeated using a 20 × 20 uniform grid and produced a less accurate (slightly oscillatory) result than the adaptive grid solution. The final adaptive grid involved fewer unknowns and the solution was more efficient. The results shown in Figures 14 and 15 agree closely with those given in Zebib *et. al.* [35] using a graded structured 62 × 54 mesh. .

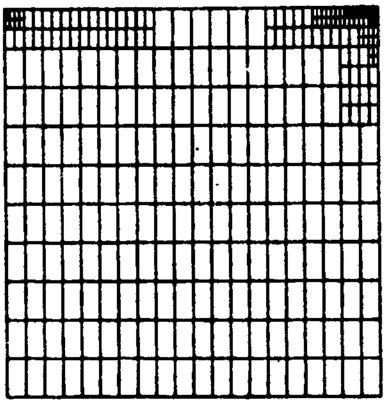


Figure 13: Adapted grid at $Re = 10000$, $Pr = 0.1$ ($Ma = 1000$) for surface tension driven flow (left wall hot, right wall cold).

5.5 Electro-Rheological (ER) Flows

Electro-rheological fluids change their material properties under the action of an electric field and exhibit shear thinning behavior similar to that of a Bingham Fluid with yield stress dependent on the E field. As a test problem to illustrate the adaptive refinement results we consider flow in a channel with fully developed inlet velocity and electrodes on opposite walls. The domain is 25mm long and the electrodes extend from $x = 5mm$ to $x = 15mm$ on each wall. The adaptive grid for for steady flow with an applied E field of $1.36MV/m$ is shown in Figure 16. There is strong shear thinning near the walls causing refinement and

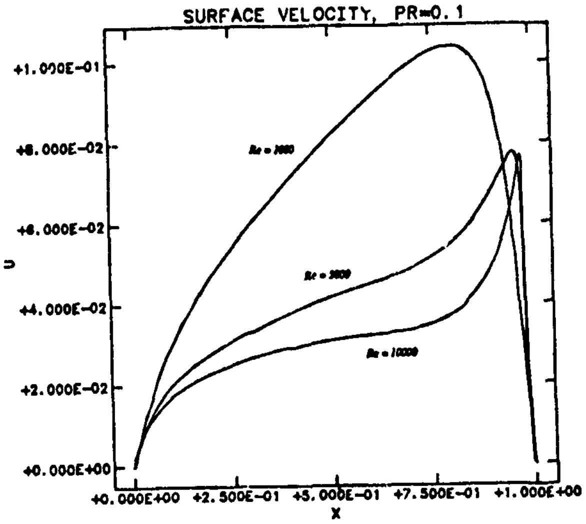


Figure 14: Surface velocity profiles for $Pr = 0.1$, $Re = 1000$, 5000 and 10000 computed from their respective adaptive grids.

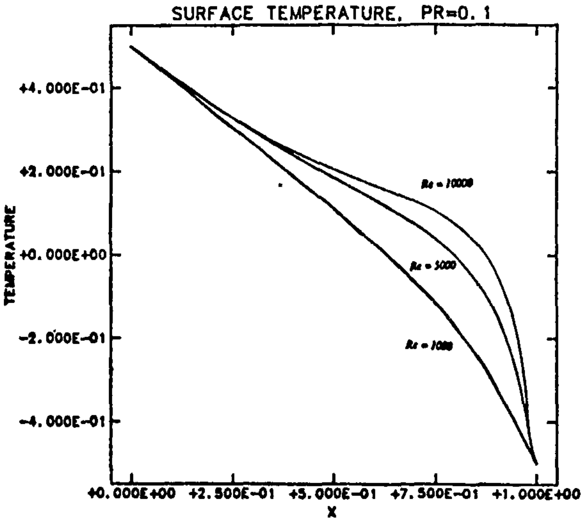


Figure 15: Surface temperature profiles for $Pr = 0.1$, $Re = 1000$, 5000 and 10000 computed on adaptive grids.

the central flow is plug like (so the grid remains unrefined here). Further details and results are given in Wang *et. al.* [34].

5.6 Refinement and Coarsening: Pulsatile Flow

Mesh coarsening (recombination or unrefinement) can be easily incorporated in the algorithm and data structure given here. Assume we have an active quartet of elements obtained at some previous refinement step. These elements can be recombined to their father element. This involves deleting the pointer from the father and releasing storage locations occupied by the quartet. These storage locations are then available for storage of other new quartets created elsewhere in the grid by refinement.

Note that at each timestep both refinement and unrefinement may be needed at different parts of the grid. In the present algorithm we first exercise unrefinement and then refinement as before. We then proceed to integrate the solution through the next timestep on this

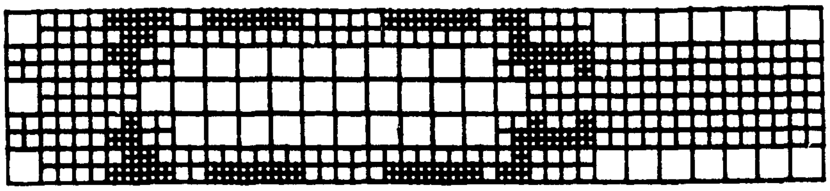


Figure 16: Mesh for *ER* problem after 2 refinements.

mesh. here a single unrefinement—refinement adjustment is made in any given timestep. If the timestep is too large more mesh adjustment within the step may be warranted. In this case the timestep solution can be repeated until the desired mesh and accuracy is achieved. Alternatively, a good timestep selection scheme such as those in ODE integrators may be devised to adjust the step adaptively.

As a test case we have considered periodic flow in a tube. The specified velocity at the inlet is sinusoidal and during the pulsed cycle, there is substantial backflow near the wall. This leads at certain stages of the periodic cycle to a turning flow in the interior. This flow behavior is reflected in the nature of the grid, which varies from refinement near the wall to refinement in this interior zone at intermediate times in the cycle. The grid and flow velocities at two different times in the period are shown in Figure 17.

6 Conclusion

An adaptive refinement and recombination scheme for steady and transient viscous flow and transport problems has been developed and implemented using biquadratic elements for the velocity and temperature fields and a bilinear basis for the pressure field. Solution of the linear Jacobian systems is achieved with a frontal solver, and the data structure is integrated into the overall solution procedure. Several numerical test cases involving Newtonian and non-Newtonian flows, including free surfaces and coupled flow and heat transfer, have been considered. These test cases demonstrate the power of the adaptive refinement scheme.

7 Acknowledgement

This research has been supported in part by the Texas Advanced Technology Research Program and by the National Science Foundation. We express our appreciation to Dr. R. McLay for his comments and suggestions.

References

[1] Babuška, I., and W.C. Rheinboldt, "Reliable Error Estimation and Mesh Adaptation for the Finite Element Method," *Computer Methods for Nonlinear Mechanics*, North Holland, New York, 1980.

[2] Babuška, I., and W.C. Rheinboldt, "Error Estimates for Adaptive Finite Element Computations," *SIAM J. Num. Anal.* **15**,736–787, 1978.

[3] Babuška, I., O.C. Zienkiewicz, J. Gago, and E.R. de A. Olivera (eds.), "Accuracy Estimates and Adaptive Refinement in Finite Element Computations," John Wiley and Sons, 1986.

[4] Bank, R.E., "Locally Computed Error Estimates for Ellilptic Equations," Proc. ARFEC, Lisbon, Portugal, Vol. 1, 21–30, 1984.

[5] Bank, R.E., and A.H. Sherman, "A Refinement Algorithm and Dynamic Data Structure for Finite Element Meshes," CNA Report TR-160, The University of Texas at Austin, 1980.

[6] Bird, R.B., R.C. Armstrong and O. Hassanger, "Dynamics of Polymetric Liquids," Vol. I, Fluid Mechanics, John Wiley and Sons, 1977.

[7] Carey, G.F., "A Mesh Refinement Scheme for Finite Element Computations," *Computer Methods in Applied Mechanics and Engineering*, **7**, 93–105, 1976.

[8] Carey, G.F. and D.L. Humphrey, "Mesh Refinement and Iterative Solution Methods for Finite Element Computations," *International Journal of Numerical Methods Engineering*, **17**, 1717–1734, 1981.

[9] Carey, G.F., and J.T. Oden, *Finite Elements: Fluid Mechanics*, Prentice Hall, N.J., 1986.

[10] Carey, G.F. and M. Seager, "Projection and Iteration in Adaptive Finite Element Refinement," *Int. J. for Num. Meth. in Eng.*, **21**, 1681–1695, 1985.

[11] Carey, G.F., M. Sharma, and K.C. Wang, "A Class of Data Structures for 2-D and 3-D Adaptive Refinement," to appear in *Int. J. Numer. Methods Eng.*, 1989.

[12] Carey, G.F., M. Sharma, K.C. Wang, and A. Pardhanani, "Some Aspects of Adaptive Grid Computations," Computers and Structures, 1988.

[13] Diaz, J.C., R.E. Ewing, R.W. Jones, A.E. Macdonald, L.M. Uhler, and D.U. von Rosenberg, "Self-Adaptive Local Grid Refinement for Time Dependent Two-Dimensional Simulation," in *Finite Elements in Fluids*, R.H. Gallagher *et al.* (eds.) Vol. VI, 279–290, John Wiley and Sons, 1986.

[14] Devloo, P., J.T. Oden, and T. Strouboulis, "Implementation of an Adaptive Refinement Technique for the SUPG Algorithm," *Computer Methods in Applied Mechanics and Engineering*, **61**, 339–358, 1987.

[15] Gartling, D.K., "Finite Element Methods for Non-Newtonian Flows," Sandia Report, SAND85-1703, 1986.

[16] George, A., "Nested Dissection of a Regular Finite Element Mesh," *SIAM J. Num. Anal.*, **10**, 406–430, 1973.

[17] Ghia, K. and V. Ghia, "Advances in Grid Generation," *ASME Monograph*, FGD-5, 1983.

[18] Gray, D.D. and A. Giorgini, "The Validity of the Boussinesq Approximation for Liquids and Gases," *Int. J. Heat Mass Transfer*, Vol. 19, 545–551, 1976.

[19] Jiang, B.N. and G.F. Carey, "Adaptive Refinement for Least Square Finite Elements with Element-by-Element Conjugate Gradient Solution," *Int. J. for Num. Methods in Eng.*, **24**, 569–580, 1987.

[20] Lohner, R., K. Morgan, and O.C. Zienkiewicz, "An Adaptive Finite Element Procedure for High speed Flows," *Computer Methods in Applied Mechanics and Engineering*, **51**, 441–465, 1985.

[21] Ludwig, R.A.. J.E. Flaherty, F. Guerinoni, P.L. Baehmann and M.S. Shepard, "Adaptive solutions of the Euler Equations Using Finite Quadtree and Octrec Grids," *Computers and Structures*, **30**, 327–336, 1988.

[22] McCormick, S., "Multigrid Methods for Variational Problems: Further Results," *SIAM J. Num. Anal.* **1**, 255–263, 1984.

[23] McLay, R.T., "Finite Element Simulation of Coupled Fluid Flow, Heat Transfer and Magnetic Fields with Applications to Welding," Ph.D. Dissertation, Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, August, 1988.

[24] McLay, R. T., and G. F. Carey, "Coupled Heat Transfer, Viscous Flow and Magnetic Effects in Weld Pool Analysis," *Int. J. Numer. Meth. Fluids*, **9**, 713–730, 1989.

[25] Ostrach, S., "Low-Gravity Fluid Flows," *Ann. Rev. Fluid Mech.*, **14**, 313, 1982.

[26] Patankar, S. V., "A Calculation Procedure for Two-Dimensional Elliptic Situations," *Numer Heat Trans.*, **4**, 409, 1981.

[27] Rank, E., "An Adaptive HP-Version in Finite Element Method," in *Numerical Techniques for Eng. Analysis and Design*, edited by G.N. Pande and J. Middleton, Martinus Nijhoff Publishers, 1987.

[28] Rheinboldt, W.C. and C.K. Mesztenyi, "On a Data Structure for Adaptive Finite Element Mesh Refinement," *ACM Trans. Math Soft.*, Vol. 6, No. 2, 166–187, 1980.

[29] Rivara, M.C., "Design and Data Structure of Fully Adaptive, Multigrid, Finite Element Software," *ACM Trans. Math. Soft.*, Vol. 10, No. 3, 242–264, 1984.

[30] Sharma, M. and G.F. Carey, "Adaptive Refinement and Iterative Solution in Semiconductor Device Simulation," *Proc of MCC-University Research Symposium*, July 1987.

[31] Shepard, M. and R.H. Gallagher, "Finite Element Grid Optimization," *ASME Monograph*, 1980.

[32] Strani, M., R. Piva, and G. Graziania, "Thermocapillary Convection in a Rectangular Cavity: Asymptotic Theory and Numerical Simulation," *J. Fluid Mech.* **130**, 347, 1983.

[33] Szabo, B., "Estimation and Control of Error Based on P-Convergence," in *Accuracy Estimates and Adaptive Refinements in Finite Element Computation*, Babuška, I., *et al.* (eds.), John Wiley and Sons, 1986.

[34] Wang, K. C., R. Mclay, and G. F. Carey, "ER Fluid Modelling," Proc. 2nd International Conference for ER Fluids, Technomic Pub., Lancaster, PA, 1990 (to appear).

[35] Zebib, A., G. M. Homsy, and E. Meiburg, "High Manangoni Number Convection in a Square Cavity," *Phys. Fluids*, 28,12, 3467–3476, 1985.

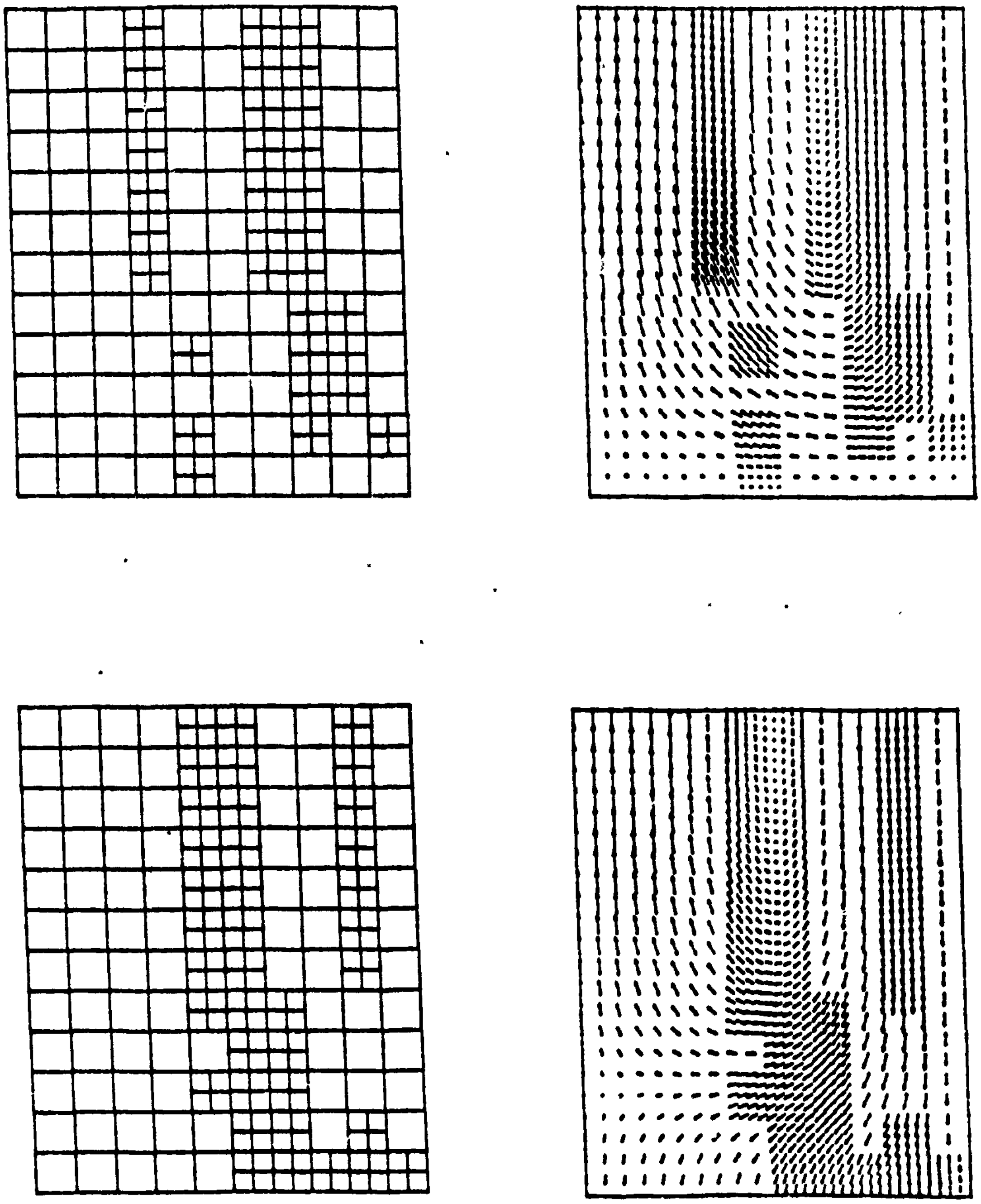


Figure 17: Grids and velocity vectors at (a) $\omega t = \pi$ and (b) $\omega t = \pi + \Delta t$.

On the Use of Hierarchical Models in Engineering Analysis

by

Klaus-Jürgen Bathe Nam-Sua Lee Miguel Luiz Bucalem
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We discuss the use of hierarchical mathematical models in finite element analysis. The use of such models is demonstrated in the analysis of a simply-supported plate and the analysis of a folded plate/shell structure. As the mathematical models are refined, new phenomena are predicted (such as boundary layers) that need careful interpretation. We conclude that the use of hierarchical models can be an important ingredient of a reliable engineering analysis.

J. Computer Methods in Applied Mechanics and Engineering, in press

1. Introduction

Finite element methods are already used very widely in engineering analysis and we can expect that this use will still increase significantly over the years to come. The issues that arise in the proper and reliable modeling by finite elements are therefore of much concern.

Figure 1 summarizes the process of finite element analysis [1] . The physical problem would typically be an actual structure or structural component subjected to certain loads. The idealization of the physical problem to a mathematical (or mechanical) model involves certain assumptions that together lead to differential equations governing the mathematical model. The finite element analysis solves this model. Since the finite element solution technique is a numerical procedure, it is necessary to assess the solution accuracy. If the accuracy criteria are not met, the numerical (i.e. finite element) solution has to be repeated with refined solution parameters (such as finer meshes or higher-order elements) until a sufficient accuracy is reached.

It is clear that the finite element solution will only solve the mathematical model, and the assumptions in this model will be reflected in the predicted response. We cannot expect any more information in the prediction of the physical phenomena than the information contained in the mathematical model. Hence, the choice of an appropriate mathematical model is crucial and completely decides the insight into the actual physical problem that we can obtain by the analysis.

Let us note here also that, by our analysis, we can of course only obtain insight into the physical problem considered: we cannot predict the response of the physical problem exactly, because it is impossible to reproduce even in the most refined mathematical model all the information that is contained in the physical problem.

Once a mathematical model has been solved accurately and the results have been interpreted we may well decide to consider next a refined mathematical model in order to increase our insight into the response of the physical problem. Furthermore, a change in the physical problem may be necessary and this in turn will also lead to additional mathematical models and finite element solutions, see Fig. 1.

The key step in engineering analysis is therefore the choice of the appropriate mathematical models. These will clearly be selected depending on what phenomena are to be predicted, and it is most important to select mathematical models that are reliable and effective in predicting the quantities sought. To define reliability and effectiveness of a chosen model we think of a very-comprehensive mathematical model of the physical problem and measure the response of our chosen model against the response of that comprehensive model. In general, the very-comprehensive mathematical model will be a fully 3-D description that also includes nonlinear effects. The most effective mathematical model for the analysis is then surely that one which yields the required response to a sufficient accuracy and at least cost. The chosen mathematical model is reliable if the required response is known to be predicted within a selected level of accuracy measured on the response of the very-comprehensive mathematical model. Hence to assess the results obtained by the solution of a chosen mathematical model, it may be necessary to also solve higher-order mathematical models.

The above considerations lead us to the notion and use of hierarchical models : a sequence of mathematical models that include increasingly more complex effects. For example, a beam structure (using engineering terminology) may first be analyzed using Bernoulli beam theory, then Timoshenko beam theory, then 2-D plane stress theory and finally using a fully 3-D continuum model and in each case nonlinear effects may be included. Clearly, with this set of hierarchical models the analysis will include ever more complex response effects, but also lead to increasingly more costly solutions. As is well-known, a fully 3-D analysis is about an order of magnitude more expensive (in computer and man-time costs) than a 2-D solution.

In this paper we focus our attention onto some of the new phenomena that may be represented by increasing the complexity of the mathematical model. As we shall demonstrate, the results obtained with each model must be carefully interpreted and can yield some surprises.

Our objective in this paper is to demonstrate the use of hierarchical models in two examples : the analysis of a simply supported plate and the analysis of a folded plate/shell structure. Both analysis problems appear to be rather simple problems; however, as we shall show, there are some most interesting phenomena that appear as the mathematical models are refined.

In Section 2 of the paper we consider the analysis of the plate using the Kirchhoff and Reissner/Mindlin plate theories and in Section 3 we present the analysis of the folded plate/shell structure using beam, shell and fully 3-D continuum mechanics theories. In each case we focus our attention onto the mathematical models and the response predicted with each model.

For the low-order mathematical models very accurate finite element solutions can be obtained without much difficulty (which we accept as the exact solutions of the mathematical models), but as fully 3-D mathematical models are considered – and hence very detailed effects are sought in the analysis – the solution of the mathematical model would in practice only be obtained to a certain level of accuracy. This practical limitation necessitates an additional consideration; namely that the predicted response be interpreted using also the level of accuracy attained in the solution of the mathematical model. However, in particular, this practical limitation also points out that care must be exercised in the choice of the mathematical models not to introduce idealizations which result into solution difficulties for phenomena that are not present in the very-comprehensive mathematical model (which is our most accurate model of the physical problem).

We conclude our paper in Section 4, where we summarize that the use of hierarchical (mathematical) models can be an important ingredient of a reliable engineering analysis.

2. Analysis of a simply-supported plate using the Kirchhoff and Reissner/Mindlin plate theories

We consider the analysis of a simply-supported square plate subjected to a distributed transverse load p per unit area. The plate has thickness h and side-length L, and is shown in Fig. 2.

In Kirchhoff plate theory the following equations govern the response of the mathematical model of the plate [2] ,

$$\nabla^4 w = \frac{p}{D} \ \mathit{in}\ \mathcal{A} \tag{1}$$

where w is the transverse displacement of the plate, and D is the flexural rigidity

$$D = \frac{Eh^3}{12(1-\nu^2)} \tag{2}$$

with E and ν the Young's modulus and Poisson ratio, respectively.

The boundary conditions for the plate are

$$\left.\begin{array}{r} w = 0 \\ M_n = 0 \end{array}\right\} \ \mathit{on}\ \Gamma \tag{3}$$

where M_n is the moment normal to the edge of the plate (corresponding to a vector along the edge of the plate). In the above equations, $\mathcal{A}$ is the mid-surface domain of the plate and Γ denotes the edges of the plate.

The detailed solution of this mathematical model of the plate is given in many textbooks (e.g. [2]). The solution directly results into the evaluation of the transverse displacement w, plate moments and shear forces in the domain $\mathcal{A}$. However, considering the force and moment conditions at the plate edges, i.e. at the boundary Γ, the twisting moment need to be converted to contribute to the edge transverse shear force. This "post-processing" procedure yields distributed shear forces along the edges of the plate plus concentrated forces at the corners.

The above briefly described Kirchhoff plate model is used very widely in engineering analysis. The model is simple to use in some respects because w is the only kinematic variable, and difficult to use in other respects. Namely, modeling difficulties arise when the constraints of the model are too severe for the physical situation considered. Specifically, these constraints do not allow for shear deformations and cause corner singularities that can have a severe and paradoxical effect on the predicted response of the plate [3] (Refer here also to the Babuška paradox in the analysis of a circular plate [4]).

However, considering finite element analysis the major difficulties in the use of the Kirchhoff plate model pertain to the continuity requirements on w and its derivatives over the element boundaries [1].

For the above reasons, the use of the Reissner/Mindlin plate theory [5, 6] for the analysis of arbitrary plates has been given much attention during the recent years. This mathematical model, when compared to the Kirchhoff plate model, is hierarchically a higher-order model and in the following we want to compare some results obtained with the Reissner/Mindlin model to those of the Kirchhoff model.

Considering the equations governing the response of the simply-supported plate the Reissner/Mindlin mathematical model can be written as [3]

$$\nabla^4 w = \frac{p}{D} - \frac{h^2}{6k(1-\nu)} \frac{\nabla^2 p}{D} \tag{4}$$

$$\frac{h^2}{12k} \nabla^2 \Omega - \Omega = 0 \tag{5}$$

where Ω is the in-plane twist

$$\Omega = \frac{1}{2}(\frac{\partial \theta_y}{\partial x} - \frac{\partial \theta_x}{\partial_y}) \tag{6}$$

In Eqs. (4) to (6), w is the transverse displacement of the mid-surface of the plate, h is the plate thickness, k is the shear correction factor and θ_x and θ_y are the plate section rotations, about the y- and minus x- axes, respectively.

We note that the above governing equations allow for transverse shear deformations, and that as $k \rightarrow \infty$ (with h fixed) the Kirchhoff plate model equations are recovered, i.e. Eq. (4) becomes Eq.(1) and from Eq. (5) we obtain $\Omega = 0$. However, when using the Reissner/Mindlin plate model in engineering practice, h is usually small (i.e. $\frac{1}{1,000} \leq \frac{h}{L} \leq \frac{1}{10}$) and k is of order

1 (typically, $k = \frac{5}{6}$). The solution of Eqs. (4) and (5) then corresponds to an interior solution plus possible boundary layer corrections. The strengths of the boundary layers depend on the boundary conditions , i.e. whether simply-supported, clamped, and so on, conditions are modeled [3, 7].

The boundary conditions on w and Ω are given in general by the conditions on the kinematic variables w ,θ_x and θ_y, and on the static variables, i.e. the moment and shear forces.

For our case of the simply-supported plate we can choose between two sets of boundary conditions which are called "soft" and "hard" conditions. The choice of which of these conditions to use is of course decided by the actual physical situation to be modeled.

The soft boundary conditions are :

$$\left.\begin{array}{r} w = 0 \\ M_{nn} = 0 \\ M_{ns} = 0 \end{array}\right\} \; on\ \Gamma \tag{7}$$

and the hard boundary conditions are :

$$\left.\begin{array}{r} w = 0 \\ M_{nn} = 0 \\ \theta_s = 0 \end{array}\right\} \; on\ \Gamma \tag{8}$$

where the normal bending moment on the edge is

$$\begin{aligned} M_{nn} \;=\; & -D(\frac{\partial^2 w}{\partial n^2} + \nu\frac{\partial^2 w}{\partial s^2}) - \frac{h^2}{6k}\frac{p}{(1-\nu)} \\ & + \frac{h^2}{6k}\frac{\partial^2}{\partial s^2}(D\nabla^2 w + \frac{h^2}{6k}\frac{p}{(1-\nu)}) - \frac{h^2}{6k}D(1-\nu)\frac{\partial^2\Omega}{\partial s\partial n} \end{aligned} \tag{9}$$

and the twisting moment is

$$\begin{aligned} M_{ns} &= \text{-}D(1-\nu)\frac{\partial^2 w}{\partial n \partial s} - \frac{h^2}{6k}\frac{\partial^2}{\partial n \partial s}(D\nabla^2 w + \frac{h^2}{6k}\frac{p}{(1-\nu)}) \\ &\quad -\frac{h^2}{6k}D(1-\nu)\frac{\partial^2 \Omega}{\partial s^2} + D(1-\nu)\Omega \end{aligned} \qquad (10)$$

Here (n , s) denotes the pair of directions normal and tangential to the edge considered. A detailed analysis of the response of the plate when assuming the soft and hard boundary conditions is given in ref. [3], see also ref. [7]. In the present paper we merely want to show some results obtained for our simply - supported plate and compare these with the solution obtained using Kirchhoff plate theory.

The most interesting results are those calculated for the transverse shear forces, and we concentrate on these results in this paper. Figure 3 shows the predicted plate shear forces along the edge of the plate and Figure 4 shows the shear forces along a line near the centre of the plate. In Fig. 4 we only give the results for the plate with the $\frac{h}{L}$ ratio equal to $\frac{1}{10}$ and $\frac{1}{100}$, because the results for $\frac{h}{L} = \frac{1}{1000}$ show on the scale used only a difference right at the edge. In each case we give the solutions corresponding to the "soft" and the "hard" boundary conditions . These results show for the solutions of the Reissner/Mindlin models :

- The solutions obtained with the soft boundary conditions show boundary layers, and the strengths of these boundary layers increase as h/L

decreases.

- The solutions obtained with the hard boundary conditions do not show boundary layers and correspond to the Kirchhoff model solutions except that the increase in the edge shear force and the concentrated forces at the corners – that are due to the allocation of the twisting moment effect in Kirchhoff theory – are not directly predicted. (However, these forces could be obtained by a "post-processing" of the results for the twisting moment along the plate edges).
- Along the plate edges, the solutions for the shear force T_n obtained with the soft boundary conditions with the Reissner/Mindlin model, converge to the "corrected" Kirchhoff model solutions (corrected by allocating the twisting moment effect into the transverse shear force).

Based on the above results we can already conclude that the Reissner/Mindlin model is significantly more powerful in predicting the response of the plate. However, we also observe that the " correct " boundary conditions (corresponding to the actual physical situation to be modeled) must be chosen. Frequently it is appropriate to use the soft boundary conditions because the twisting moment M_{ns} is zero rather than the kinematic variable ϑ_s.

Hierarchically, the Reissner/Mindlin plate model includes the Kirchhoff model and can be a more reliable model to use, in particular, when shear stresses along the plate edges are to be predicted, or when plates of arbitrary geometry are analyzed [3]. We finally should mention that, of course, plate models of higher-order could also be used in the analysis of our plate or a fully 3-D solution could be sought [1, 4, 8]. When using such higher-order models, additional variables are introduced that can represent additional phenomena and hence can lead to further insight into the behavior of the actual physical plate considered.

3. Analysis of a folded plate/shell structure

We next consider the structure shown in Fig. 5. This structure can be analyzed using Bernoulli beam theory, Timoshenko beam theory, 2-D plane stress theory, shell theory and a fully 3-D continuum theory. It is clear that as we consider a more complex model the assumptions on the kinematic and static behavior of the mathematical model are different than for the lower-order models and the appropriate boundary conditions must be identified. As in the analysis of the plate (see Section 2), the choice of boundary conditions may affect the solution results significantly.

Our objective below is to simply state the mathematical models we have used for the analysis of the structure and give some solution results. The response predictions demonstrate that, as expected, certain solution parameters can only be predicted with a sufficiently high-order model. However, the results and discussion will also show that as the order of the mathematical model is increased, the solution effort (by finite element analysis) can become very large. In such case, it is only practical to solve for the response of the mathematical model to a certain level of accuracy and include the accuracy considerations in the interpretation of the results.

Hence in this section we shall also describe briefly the finite element discretizations used to solve the mathematical models.

3.1 Timoshenko beam model

The simplest mathematical model to represent the folded plate/shell structure is a beam model based on Bernoulli or Timoshenko beam theory. We considered a Timoshenko beam model and solved for the exact response of the mathematical model by using one cubic-isoparametric beam element [1] each, for the vertical plate, and for the horizontal plate (with symmetry conditions at the mid-section of the structure).

3.2 Shell model

The second mathematical model considered for the analysis of the folded plate/shell structure was based on the Reissner/Mindlin plate theory to model the plate actions, as briefly summarized in Section 2, and the usual 2-D plane stress theory to model the membrane in-plane actions. The solution to this mathematical model was obtained using the 16-node displacement-based shell elements described in detail in refs. [1,9].

Figures 6 to 8 show the mathematical model and the finite element mesh used with the 16-node shell elements. The finite element mesh was deemed fine enough because the stress jumps at the nodal points were negligibly small and smooth stress variations were predicted (Of course, an accurate error indicator would be based on the residual of the differential equations of equilibrium [10]). Figure 6 also shows the modeling used at the junction of the horizontal and vertical plates. As described in Section 2, the mathematical model can here be subjected to hard boundary conditions ($\theta_y = \theta_z = 0$) or soft boundary conditions (θ_y free and θ_z free). We obtained our analysis results, given in section 3.4, using soft boundary conditions.

3.3 3-D model

The third mathematical model was constructed to capture the three-dimensional effects at the junction of the plates. Hence we used 3-D continuum mechanics theory at the junction of the plates and the assumptions of the shell model discussed in the previous section at a distance from that junction, see Fig. 9. Of course, a coupling of the 3-D continuum mechanics assumptions and shell theory assumptions was necessary at two sections of the plates as shown in Fig. 9. We refer in this paper to this model as the "3-D model".

For the solution of the mathematical model, 960 three-dimensional 20-node elements were used to represent the 3-D part of the mathematical model, and 300 sixteen-node shell elements were used to represent the part mathematically modeled by shell theory. Figure 10 shows a section of the finite element discretization and indicates the element grading used.

Two considerations may be mentioned. Firstly, of course, a 3-D mathematical model could havc been used for the complete structure. This would have required much more solution effort, and since we are only interested in studying the behavior at the junction of the plates, a fully 3-D model was not necessary.

Secondly, the finite element solution of the mathematical model was only sought to a certain level of accuracy. Our objective was to solve the mathematical model so as to identify stress variations that occur over at least a distance of $\frac{1}{10}$ *th* the thickness of the plates. Hence, our finite element discretization was not intended to solve for stress variations that are described in the mathematical model but that only occur over extremely small distances.

Figure 11 shows a pressure band plot calculated with our finite element discretization. The bands are quite smooth which indicates that our finite element mesh is adequate (for the objective stated above) [11].

3.4 Solution results

Our aim in this section is to show some results obtained with the three different models for the folded plate/shell structure. We want to compare these results and briefly discuss them with some emphasis on the hierarchical nature of the models.

The transverse displacement at the section of the load application is quite accurately predicted using either of the three models, see Fig. 12. Of course, the beam model gives a constant z-displacement as a function of x, whereas the shell and 3-D models predict a larger transverse displacement at the plate edges than in the centre. Notice that, as expected the 3-D model gives the smallest transverse displacement and the beam model gives the largest displacement.

Hence, considering our discussion in Section 1, the beam model displacement results are reliable if the 3-D model is considered the very-comprehensive model and the selected level of accuracy for the displacement under the load is ten percent on the response of that model.

Figures 13 and 14 show the predicted longitudinal stresses using the three models. As probably expected, except for the region near the junction of the plates, all three models predict closely the same stress distributions. However, at the junction of the plates we observe the following :

(a) On the top surface and along the centre line (i.e. along the line OPQ in Fig. 5), the beam and shell model predictions are close and only the 3-D model can of course represent the zero stress conditions on the free surfaces, see Fig. 13a.

(b) On the top surface and along the free edge (i.e. along line ABC), the shell model predicts a longitudinal stress in-between the results of the

3-D model and the beam model, see Fig. 13b.

(c) On the bottom surface and along the centre line, only the 3-D model shows a sudden stress rise (due to the stress singularity), see Fig. 14a.

(d) On the bottom surface and along the free edge, the shell and 3-D models predict a stress drop, see Fig. 14b.

The results quoted in (a) and (c) are quite expected, whereas the shell model results referred to in (b) may be unexpected, and the shell and 3-D model results quoted in (d) may well be a surprise. Namely, we would expect that the sharp corner causes a sudden stress rise instead of the decrease in the stress. However, these results are explained in that the finite element solution of the mathematical model is simply not accurate enough to show the still possible sudden stress rise (see Section 3.3). Indeed, Babuška has solved this mathematical model accurately (with a huge computational effort) and has shown that this stress rise still occurs at about a distance of $\frac{1}{1000}$ times the thickness of the plate from the corner [4]. While the results of Babuška are most valuable, and point out some very interesting aspects of this particular mathematical model, a more comprehensive model of the actual physical problem (including the actual geometry at the corner and nonlinear effects) would of course not show such a stress distribution.

4. Concluding Remarks

Our objective in this paper was to discuss certain key aspects of the finite element analysis process:

- The crucial step of a finite element analysis is always the selection of an appropriate mathematical model of the physical problem.
- The choice of the mathematical model depends on the effects and quantities to be predicted.
- The mathematical model is to be effective and reliable for the prediction (with "effective" and "reliable" defined in the paper).
- The finite element solution of the chosen mathematical model should be obtained to a level of accuracy measured against the response to be expected from a very-comprehensive mathematical model.

Good engineering analysis is of course an art and is usually based on a great deal of experience. However, the considerations given in this paper point out that whichever approach of analysis is followed, the use of hierarchical mathematical models can be of significant value: with the use of hierarchical mathematical models the analysis process has structure and leads to results that can be accepted with confidence. Although not considered in this paper, the use of hierarchical models is most important when there is need for a nonlinear response prediction [1].

References

[1] K. J. Bathe, Finite Element Procedures in Engineering Analysis, Prentice-Hall, Englewood Cliffs, New Jersey (1982).

[2] S. P. Timoshenko and S. Woinowsky-Krieger, Theory of Plates and Shells, McGraw-Hill (1959).

[3] B. Häggblad and K. J. Bathe, "Specifications of Boundary Conditions for Reissner/Mindlin Plate Bending Finite Elements", Int. J. Num. Meth. in Eng., in press.

[4] I. Babuška "The Problem of Modeling Elastomechanics in Engineering", J. Computer Methods in Applied Mechanics and Engineering, in press.

[5] E. Reissner, "The Effect of Transverse Shear Deformation on the Bending of Elastic Plates", J. Appl. Mech., 12, A69 (1945).

[6] R. D. Mindlin, "Influence of Rotary Inertia and Shear on Flexural Motions of Isotropic Elastic Plates", J. Appl. Mech., 18, 31-38 (1951).

[7] D. N. Arnold and R. S. Falk, "Edge Effects in the Reissner-Mindlin Plate Theory", in Analytical and Computational Models of Shells, (A. K. Noor, et. al., Eds.), ASME Special Publication, 71-89 (1989).

[8] I. Babuška and T. Scapolla,"Benchmark Computation and Performance Evaluation for a Rhombic Plate Bending Problem", Int. J. Num. Meth. in Eng., Vol. 28, 155-178 (1989).

[9] K. J. Bathe and S. Bolourchi, " A Geometric and Material Nonlinear Plate and Shell Element" J. Comput. Struct., 11, 23-48 (1979)

[10] S. W. Chae and K. J. Bathe, "On Automatic Mesh Construction and Mesh Refinement in Finite Element Analysis", J. Comput. Struct., 32, No 3/4, 911-936 (1989)

[11] T. Sussman and K. J. Bathe, "Studies of Finite Element Procedures — Stress Band Plots and the Evaluation of Finite Element Meshes", Eng. Computations, 3, 178-191 (1986)

A DISCOURSE ON THE STABILITY CONDITIONS FOR MIXED FINITE ELEMENT FORMULATIONS

by

Franco Brezzi
Dipartimento di Meccanica Strutturale and Istituto di Analisi
Numerica del C.N.R., 27100 Pavia, Italy

and

Klaus-Jürgen Bathe
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

We discuss the general mathematical conditions for solvability, stability and optimal error bounds of mixed finite element discretizations. Our objective is to present these conditions with relatively simple arguments. We present the conditions for solvability and stability by considering the general coefficient matrix of mixed finite element discretizations, and then deduce the conditions for optimal error bounds for the distance between the finite element solutions and the exact solution of the mathematical problem. To exemplify our presentation we consider the solutions of various example problems. Finally, we also present a numerical test that is useful to identify numerically whether, for the solution of the general Stokes flow problem, a given finite element discretization satisfies the stability and optimal error bound conditions.

1. INTRODUCTION

During the recent years it has been recognized to an increasing extent that the use of mixed finite elements can be of great benefit and may even be necessary to obtain reliable and accurate solutions in certain fields of engineering analysis. Mixed finite elements are currently used with much success in the solution of incompressible fluid flows, and continue to provide great promise for the analysis of solids and structures [1,2].

Of course, the largest area of finite element applications is still structural analysis and mixed finite elements are, in principle, much suited for use in the analysis of almost incompressible media (for example, for the analysis of rubber-like materials, elasto-plasticity and creep) and the analysis of plates and shells. However, although many mixed finite elements have been proposed over the last two decades in the research literature, it is apparent that mixed finite elements are hardly used in practical structural analysis.

The reason why mixed finite elements are not used abundantly in engineering practice is that their predictive behavior is much more difficult to assess than for the conventional and commonly used displacement-based elements. Whereas displacement-based elements, once formulated and shown to work well on certain sets of examples (including the patch tests), can be generally employed, mixed finite elements cannot be recommended for general use unless a deeper analysis and understanding is available. Namely, considering a certain category of problems, a mixed finite element may work well in the solution of certain problems but perform very poorly on other problems. Therefore, a mathematical analysis (even a limited one) for the stability and convergence of a proposed formulation is an important requirement. Such mathematical analysis should give sufficient insight as to the general applicability of the finite element under consideration, and is in general no easy task.

Some researchers have proposed some easily applied "counting rules" to assess whether a mixed finite element can be recommended [3,4]. However, such rules can at best give some guide lines, and do not give the necessary information to assess whether an element is stable and accurate.

Considering mixed finite element discretizations, we recognize that they are governed by a system of equations with a coefficient matrix, C, that we may write as

$$C = \begin{bmatrix} A & B^t \\ B & 0 \end{bmatrix} \tag{1.1}$$

We quote as a main example the analysis of incompressible fluid flow, the Stokes problem, when using the velocity-pressure formulation. Other important examples of interest are the analysis of incompressible solids and the analysis of plates and shells. In principle, many solutions can be formulated using a mixed or hybrid method that results into the coefficient matrix (1.1), because this matrix is reached by minimizing a functional under linear constraints [1].

The general mathematical theory for the solution of problems that are governed by the coefficient matrix in (1.1) is now quite well established and the detailed applications of this theory to a number of important problem categories is available. We know necessary and sufficient conditions for the existence and uniqueness of the solution, both for the continuous and the discretized problems. We also know necessary and sufficient conditions on the choice of the discretizations in order to have optimal error bounds [1,5]. This information is most valuable for the design *and* analysis of mixed finite elements because the basic mathematical results are quite generally applicable (while the detailed application to problem areas may of course not be straight-forward).

Our objective in this paper is two-fold. The first aim is to present the general mathematical results quoted above with relatively simple arguments. For this purpose we consider the general coefficient matrix of mixed finite element formulations and deduce the conditions of solvability and stability. In proceeding this way, we refer to the continuous problem only when necessary (since the treatment of the continuous problem requires a background in functional analysis) and we concentrate on the discretized (finite-dimensional) problem. However, we ceed in pointing out the basic mathematical conditions on the discretization and in showing that they are necessary to have stability and optimal error estimates.

Our second aim in this paper is to propose a simple numerical procedure for checking whether the above mathematical conditions are satisfied for a given mixed finite element formulation. Such a procedure is useful because it may be employed to check a formulation and its computer program implementation (much like the patch test is used for incompatible displacement-based finite element formulations). We consider in this discussion the analysis of incompressible fluid flow and our test is closely related to "Fortin's trick" to identify whether the mathematical conditions of stability and optimal error bounds are satisfied.

The paper is organized into the following sections. In Section 2 we recall some basic properties of square matrix systems and introduce the basic concepts of stability and optimality. In Section 3 we ther deal with the special case of systems of the form (1.1); hence here we focus onto the analysis of mixed finite element formulations in detail. Finally, in Section 4 we discuss two applications and introduce our test for checking the good quality of a given discretization, using as an example the case of an incompressible fluid. We then conclude our presentation in Section 5.

2. SOME PRELIMINARIES AND THE GENERAL PROBLEM OF SOLVABILITY AND STABILITY

Let us consider the general case of an $N \times N$ matrix M and the associated system

$$\left.\begin{array}{l}\text{given } b \in \mathbb{R}^N \text{ find } x \in \mathbb{R}^N \text{ such that} \\ Mx = b\end{array}\right\} \tag{2.1}$$

The following theorem is a well-known cornerstone in linear algebra.

THEOREM 2.1 Problem (2.1) has a unique solution for every given right-hand side b if, and only if, the associated homogeneous system $Mx = 0$ has only the solution $x = 0$. □

In other words, in order to have a *solvable problem* in (2.1) for every possible $b \in \mathbb{R}^N$ we need the following condition to hold:

$$\textit{if } Mx = 0, \quad \textit{then } x = 0. \tag{2.2}$$

Condition (2.2) answers the problem of the solvability of (2.1) but not of its *stability*. Roughly speaking, we would like that a small change in b determines only a small change in x. However, in order to measure the magnitude of such change we have to introduce norms. Assume that we choose a norm $\| \quad \|_L$ for measuring the size of solutions and a norm $\| \quad \|_R$ for right-hand sides. In principle, we are allowed to choose the same norm for both, but we shall see that this, in general, is not the most convenient choice. We also point out explicitly that, in finite dimensional spaces, all norms are equivalent, in the sense that, for any two norms $\| \quad \|_{S_1}$ and $\| \quad \|_{S_2}$ in $\mathbb{R}^N$ there exist two positive constants s_1 and s_2 such that

$$\| v \|_{S_1} \leq s_1 \| v \|_{S_2} \tag{2.3}$$

$$\| v \|_{S_2} \leq s_2 \| v \|_{S_1} \tag{2.4}$$

for every vector v in $\mathbb{R}^N$. However, these constants s_1 and s_2 will, in general, depend on the dimension N.

EXAMPLE . This is a very simple example, only used to fix our ideas.[2]. Let

$$\| v \|_{S_1} := \max_i |v_i| = \| v \|_{\ell_\infty} \tag{2.5}$$

$$\| v \|_{S_2} := \sum_i |v_i| = \| v \|_{\ell_1} \tag{2.6}$$

then it is easy to see that

$$\max_i |v_i| \leq \sum_i |v_i| \tag{2.7}$$

$$\sum_i |v_i| \leq N \max_i |v_i| \tag{2.8}$$

so that $s_1 = 1$ and $s_2 = N$. Similarly we have for the Euclidean norm

$$\| v \|_E := (\sum_i |v_i|^2)^{1/2} = \| v \|_{\ell_2}, \tag{2.9}$$

that

$$\| v \|_{\ell_1} \leq \sqrt{N} \| v \|_{\ell_2} \; ; \quad \| v \|_{\ell_2} \leq \sqrt{N} \| v \|_{\ell_\infty} \tag{2.10}$$

□

We have seen that the choice of one norm or another can change, asymptotically, the dependence on N of the various constants. We shall come back to this point with useful guidelines for the most convenient choices. For the moment, we assume that the choice of $\| \quad \|_L$ and $\| \quad \|_R$ has been performed and define stability in terms of these norms.

DEFINITION Let M be a non-singular $N \times N$ matrix. We define the stability constant of M with respect to the norms $\| \quad \|_L$ and $\| \quad \|_R$ as the smallest possible constant S_{LR} such that

$$\frac{\| \delta x \|_L}{\| x \|_L} \leq S_{LR} \frac{\| \delta b \|_R}{\| b \|_R} \tag{2.11}$$

for all vectors x and δx in $I\!R^N$ with $Mx =: b$ and $M\delta x =: \delta b$. □

In other words, (2.11) bounds the relative change in x (in the norm L) by means of the relative change in the right-hand side b (in the norm R). We point out that such a constant S_{LR} always exists. However, if we consider a *sequence* of problems of type (2.1) with increasing dimension N (corresponding, in general, to a finer and finer finite element mesh) we might find that the corresponding constants S_{LR} depend on N and become infinitely large when $N \to +\infty$. Thus we might say that *a sequence of problems* of the type (2.1) *is stable* with respect to the norms $\| \quad \|_L$ and $\| \quad \|_R$ if the stability constant S_{LR} is *uniformly bounded.*

We would like now to present stability from a slightly different point of view. For this, let us introduce the matrix norms

$$\| M \|_{LR} = \sup_y \frac{\| My \|_R}{\| y \|_L} \tag{2.12}$$

and

$$\| M^{-1} \|_{RL} = \sup_{z} \frac{\| M^{-1}z \|_L}{\| z \|_R}. \tag{2.13}$$

From (2.13) for $z = \delta b$ (so that $M^{-1}z = \delta x$) we easily obtain

$$\| M^{-1} \|_{RL} \geq \frac{\| \delta x \|_L}{\| \delta b \|_R} \tag{2.14}$$

while (2.12), for $y = x$ (and $My = b$) gives

$$\| M \|_{LR} \geq \frac{\| b \|_R}{\| x \|_L}. \tag{2.15}$$

From (2.14) and (2.15) we then have

$$\frac{\| \delta x \|_L}{\| x \|_L} \leq \| M \|_{LR} \cdot \| M^{-1} \|_{RL} \frac{\| \delta b \|_R}{\| b \|_R} \tag{2.16}$$

from which

$$S_{LR} = \| M \|_{LR} \;\; \| \; M^{-1} \|_{RL} \tag{2.17}$$

REMARK 2.1 Noting that, for every x, one has $x = M^{-1}Mx$ we obtain

$$\| x \|_L \leq \| M^{-1} \|_{RL} \;\; \| M \|_{LR} \;\; \| x \|_L \tag{2.18}$$

which easily implies

$$S_{LR} = \| M^{-1} \|_{RL} \;\; \| M \|_{LR} \geq 1. \tag{2.19}$$

REMARK 2.2 If we choose $\| \quad \|_L = \| \quad \|_R = \| \quad \|_E$ (Euclidean norm), and if M is symmetric and positive definite, then

$$\| M \|_{LR} = \lambda_{\max} \;\; ; \;\; \| M^{-1} \|_{LR} = 1/\lambda_{\min} \tag{2.20}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and (respectively) minimum eigenvalues of M.

Hence, for the case of M being symmetric and positive definite, we have that

$$S_{LR} = S_{EE} = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{2.21}$$

coincides with the usual *condition number*. Note however, that a different choice of norms will (obviously) produce different stability constants. For instance, by taking

$$M = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}; \| \quad \|_L = \| \quad \|_{\ell_\infty} \; ; \; \| \quad \|_R = \| \quad \|_{\ell_1} \tag{2.22}$$

(see (2.5) and (2.6) for the definition of the norms $\| \quad \|_{\ell_\infty}$ and $\| \quad \|_{\ell_1}$) we have

$$\lambda_{\max} = \frac{3+\sqrt{5}}{2}; \lambda_{\min} = \frac{3-\sqrt{5}}{2}; \| M \|_{LR} = 5; \| M^{-1} \|_{RL} = 1. \tag{2.23}$$

so that $S_{EE} = \frac{3+\sqrt{5}}{3-\sqrt{5}}$ (= condition number) while $S_{LR} = 5$. We shall see in the following that, for practical problems, we have, in a natural way, choices for the norms $\| \quad \|_L$ and $\| \quad \|_R$ for which S_{LR} will be uniformly bounded while S_{EE} is not. □

From (2.17) we see that a sequence of problems will be stable with respect to the norms $\| \quad \|_L$ and $\| \quad \|_R$ if both $\| M \|_{LR}$ and $\| M^{-1} \|_{RL}$ are uniformly bounded. In the applications it is very easy to find norms $\| \quad \|_L$ such that

$$y^t M x \leq k_M \| y \|_L \; \| x \|_L \qquad \forall x, y \tag{2.24}$$

with k_M *uniformly bounded from above and from below.* From (2.24) we have a natural choice for $\| \quad \|_R$ that produces a uniform bound for $\| M \|_{LR}$. Indeed, if we define the *dual norm* of $\| \quad \|_L$ by

$$\| z \|_{DL} := \sup_y \frac{y^t z}{\| y \|_L} \tag{2.25}$$

we have the following proposition.

PROPOSITION 2.1. Let M be an $N \times N$ matrix, let $\| \ \|_L$ be a norm in $\mathbb{R}^N$ and let k_M be the smallest possible constant for which (2.24) holds true, that is

$$k_M = \sup_{x,y} \frac{y^t M x}{\| y \|_L \ \| x \|_L}. \tag{2.26}$$

If we choose $\| \ \|_R = \| \ \|_{DL}$ (dual norm of $\| \ \|_L$ as defined in (2.25)) then

$$\| M \|_{LR} = k_M. \tag{2.27}$$

PROOF. We have

$$\begin{aligned} \| M \|_{LR} &= (\text{use } (2.12)) = \sup_x \frac{\| Mx \|_R}{\| x \|_L} = \\ &= (\text{use } \| \ \|_R = \| \ \|_{DL}) = \sup_x \frac{\| Mx \|_{DL}}{\| x \|_L} = \\ &= (\text{use } (2.25)) = \sup_x \left\{ \frac{1}{\| x \|_L} \sup_y \frac{y^t M x}{\| y \|_L} \right\} = \\ &= \sup_{x,y} \frac{y^t M x}{\| x \|_L \ \| y \|_L} = (\text{use } (2.26)) = k_M \end{aligned} \tag{2.28}$$

□

If we assume now that we are given a sequence of problems such that

$$y^t M x \leq k_M \| y \|_L \ \| x \|_L \qquad \forall x, y$$

with k_M uniformly bounded from above and from below, and if we choose $\| \ \|_R = \| \ \|_{DL}$ then the sequence of problems will be stable with respect to the norms $\| \ \|_L$ and $\| \ \|_R$

if and only if $\| M^{-1} \|_{RL}$ is uniformly bounded. In the following proposition we express $\| M^{-1} \|_{RL}$ ~~is~~ in terms of the norm $\| \quad \|_L$ alone.

PROPOSITION 2.2. Let M be a non-singular $N \times N$ matrix. Let $\| \quad \|_L$ be a norm in $\mathbb{R}^N$ and let $\| \quad \|_R$ be the dual norm of $\| \quad \|_L$ as defined in (2.25). Then

$$(\| M^{-1} \|_{RL})^{-1} = \inf_x \sup_y \frac{y^t M x}{\| y \|_L \ \| x \|_L}. \tag{2.29}$$

PROOF. We have

$$\begin{aligned}
(\| M^{-1} \|_{RL})^{-1} &= (\text{use } (2.13) = \left(\sup_z \frac{\| M^{-1} z \|_L}{\| z \|_R} \right)^{-1} = \inf_z \frac{\| z \|_R}{\| M^{-1} z \|_L} = \\
&= (\text{set } z = Mx) = \inf_x \frac{\| Mx \|_R}{\| x \|_L} = \\
&= (\text{use } \| \quad \|_R = \| \quad \|_{DL}) = \inf_x \frac{\| Mx \|_{DL}}{\| x \|_L} = \\
&= (\text{use } (2.25)) = \inf_x \left\{ \frac{1}{\| x \|_L} \sup_y \frac{y^t M x}{\| y \|_L} \right\} = \\
&= \inf_x \sup_y \frac{y^t M x}{\| x \|_L \ \| y \|_L}
\end{aligned} \tag{2.30}$$

□

The following proposition summarizes Propositions 2.1 and 2.2.

PROPOSITION 2.3. Let M be an $N \times N$ non-singular matrix, let $\| \quad \|_L$ be a norm in $\mathbb{R}^N$ and let $\| \quad \|_R$ be its dual norm as defined in (2.25). Setting

$$k_M = \sup_{x,y} \frac{y^t M x}{\| y \|_L \ \| x \|_L} \tag{2.31}$$

$$\gamma_M = \inf_x \sup_y \frac{y^t M x}{\| x \|_L \ \| y \|_L} \tag{2.32}$$

the stability constant S_{LR} of M is given by

$$S_{LR} = k_M/\gamma_M. \tag{2.33}$$

The proof is obvious from (2.17), (2.27), (2.29), (2.31) and (2.33).

REMARK 2.3 If we assume to be dealing with a sequence of problems where

$$y^t M x \leq k \parallel y \parallel_L \ \parallel x \parallel_L \qquad \forall\, x, y \tag{2.34}$$

with k uniformly bounded *from above*, then $k_M \leq k$ and in order to have a uniform bound for S_{LR} we only need γ_M to be uniformly bounded *from below*, that is, we need a constant $\gamma > 0$ such that

$$\inf_x \sup_y \frac{y^t M x}{\parallel x \parallel_L \ \parallel y \parallel_L} \geq \gamma > 0 \tag{2.35}$$

for every problem of the sequence.

REMARK 2.4 We remember that the solvability of (2.1) was expressed in (2.2) by

$$Mx = 0 \Rightarrow x = 0. \tag{2.36}$$

Under the assumption (2.34) we have now that the stability can be expressed by (2.35) which, in its turn, can be written as

$$\exists \gamma > 0 \text{ such that } \parallel Mx \parallel_{DL} \geq \gamma \parallel x \parallel_L \qquad \forall x. \tag{2.37}$$

Indeed (2.35) can be written as

$$\exists \gamma > 0 \text{ such that } \sup_y \frac{y^t M x}{\parallel y \parallel_L} \geq \gamma \parallel x \parallel_L \qquad \forall x, \tag{2.38}$$

which becomes (2.37) by using the definition of the dual norm (2.25). □

We end this section by analyzing the connection of the above results with the use of Galerkin methods for the discretization of variational problems. Let us consider a general linear elasticity problem characterized by a given Hilbert space W and a bilinear form $m(\phi,\psi)$ defined on $W\times W$. Given a linear functional $\beta(\phi)$ from W to $\mathbb{R}$ we want to approximate the solution of the continuous problem

$$\begin{cases} \text{find } \phi\in W \text{ such that} \\ m(\phi,\psi)=\beta(\psi) \qquad \forall\psi\in W \end{cases} \tag{2.39}$$

by means of the sequence of finite dimensional problems

$$\begin{cases} \text{find } \phi_h\in W_h \text{ such that} \\ m(\phi_h,\psi_h)=\beta(\psi_h) \qquad \forall\psi_h\in W_h \end{cases} \tag{2.40}$$

where W_h is a sequence of finite dimensional subspaces of W_h. Let us note that (2.40) is a very general mixed formulation. However, it may help the intuition of the reader to think of a displacement-based finite element discretization, which is the easiest case.

By choosing a basis $\phi^{(1)},\cdots,\phi^{(N)}$ in W_h we can associate with every vector $\xi\in\mathbb{R}^N$ the element

$$\sum_i \xi_i\phi^{(i)}\in W_h \tag{2.41}$$

(in the usual way). Every problem (2.40) has now the form (2.1) with

$$M_{ij}:=m(\phi^{(j)},\phi^{(i)});b_i:=\beta(\phi^{(i)}). \tag{2.42}$$

If the linear form $m(\phi,\psi)$ satisfies

$$m(\phi,\psi)\le k_m \parallel \phi \parallel_W \parallel \psi \parallel_W \qquad \forall\phi,\psi\in W \tag{2.43}$$

then (2.34) will easily hold with $k = k_m$ (with k_m independent of h) if we choose

$$\| \xi \|_L := \| \sum_i \xi_i \phi^{(i)} \|_W \tag{2.44}$$

as a norm in $I\!R^N$. The stability condition (2.35) can now be written in terms of the bilinear form $m(\phi, \psi)$ and of the space W_h as

$$\inf_{\phi_h \in W_h} \sup_{\psi_h \in W_h} \frac{m(\phi_h, \psi_h)}{\| \phi_h \|_W \; \| \psi_h \|_W} \geq \gamma > 0 \tag{2.45}$$

with γ independent of h.

We point out that (2.45) on one hand implies (as we have seen) the solvability of every discrete problem (2.40). On the other hand, if (2.45) holds with γ independent of h, then one can deduce optimal error bounds for the distance between the solution ϕ of (2.39) and the solution ϕ_h of (2.40). Incidentally, we point out that (2.45), together with

$$\lim_{h \to 0} \left\{ \inf_{\psi_h \in W_h} \| \psi - \psi_h \|_W \right\} = 0 \qquad \forall \psi \in W \tag{2.49}$$

implies that (2.39) has a unique solution. We shall not report here the proof of this fact (which has basically little bearing upon our discussion), and shall instead report the proof of the optimal error bounds.

THEOREM 2.2 (Babuška [6]).

Assume that the bilinear form $m(\phi, \psi)$ and the sequence of subspaces $W_h \subset W$ satisfy (2.43), (2.45) and (2.49). Let ϕ be the solution of (2.39) and ϕ_h the solution of (2.40). Then

$$\| \phi - \phi_h \|_W \leq (1 + k_m/\gamma) \inf_{\psi_h \in W_h} \| \phi - \psi_h \|_W \,. \tag{2.50}$$

PROOF. For every $\psi_h \in W_h$ we have

$$\begin{aligned}
\gamma \parallel \psi_h - \phi_h \parallel_W &\leq \text{(use (2.45))} \overset{\leq}{\vee} \sup_{\chi_h \in W_h} \frac{m(\psi_h - \phi_h, \chi_h)}{\parallel \chi_h \parallel_W} = \\
&= \text{(add and subtract } \phi) = \\
&= \sup_{\chi_h \in W_h} \{m(\psi_h - \phi, \chi_h) - m(\phi - \phi_h, \chi_h)\} \,/ \parallel \chi_h \parallel_W \\
&= \text{(use (2.39) and (2.40))} \overset{=}{\vee} \sup_{\chi_h \in W_h} \frac{m(\psi_h - \phi, \chi_h)}{\parallel \chi_h \parallel_W} \leq \qquad (2.51) \\
&\leq \text{(use (2.43))} \overset{\leq}{\vee} \sup_{\chi_h \in W_h} \frac{k_m \parallel \psi_h - \phi \parallel_W \parallel \chi_h \parallel_W}{\parallel \chi_h \parallel_W} = \\
&= k_m \parallel \psi_h - \phi \parallel_W .
\end{aligned}$$

From (2.51) we have, using the triangle inequality:

$$\begin{aligned}
\parallel \phi_h - \phi \parallel_W &\leq \parallel \phi_h - \psi_h \parallel_W + \parallel \psi_h - \phi \parallel_W \leq \\
&\leq (k_m/\gamma) \parallel \psi_h - \phi \parallel_W + \parallel \psi_h - \phi \parallel_W = \qquad (2.52) \\
&= (1 + k_m/\gamma) \parallel \psi_h - \phi \parallel_W
\end{aligned}$$

and (2.50) follows since (2.52) holds for every $\psi_h \in W_h$.

3. SOLVABILITY AND STABILITY OF MIXED FINITE ELEMENT FORMULATIONS

We consider now a special case of (2.1). Namely we assume that the matrix M has the typical form arrived at when using a mixed finite element formulation,

$$M = \begin{pmatrix} A & B^t \\ B & 0 \end{pmatrix} \tag{3.1}$$

where A is a square $NA \times NA$ matrix and B a rectangular $NB \times NA$ matrix with (obviously) $NA+NB = N$. Accordingly we split the unknown $x = (u, p)$ with $u \in \mathbb{R}^{NA}$ and $p \in \mathbb{R}^{NB}$ and the right-hand side $b = (f, g)$ with $f \in \mathbb{R}^{NA}, g \in \mathbb{R}^{NB}$. With this notation, the linear system under examination can be written as

$$\left.\begin{array}{lcl} Au + B^t p & = & f \\ Bu & = & g \end{array}\right\} \tag{3.2}$$

The analysis of the solvability, stability and optimality of mixed formulations has been performed in [5]. However, we shall follow here the more elegant presentation of Arnold [7]. In any case, the following space is of crucial importance. We set

$$K = \{v \in \mathbb{R}^{NA}, \text{ such that } Bv = 0\} \tag{3.3}$$

(in other words $K = Ker(B)$). If NK is the dimension of K we can split $\mathbb{R}^{NA}$ as

$$\mathbb{R}^{NA} = T \oplus K \tag{3.4}$$

where T is the orthogonal of K in $\mathbb{R}^{NA}$. As a consequence of (3.4) every $v \in \mathbb{R}^{NA}$ can be spiit, in a unique way, as a sum

$$v = v_T + v_K \quad \text{with } v_T \in T, v_K \in K, \quad \text{and } v_T^t v_K = 0. \tag{3.5}$$

If NT is the dimension of T, we will obviously have $NT + NK = NA$.

Let us now assume that the system of equations with the matrix M has been established in a suitable basis so that we can write the matrix A as

$$A = \begin{pmatrix} A_{TT} & A_{TK} \\ A_{KT} & A_{KK} \end{pmatrix}. \tag{3.6}$$

The notation (3.6) implies that the choice of the basis and the ordering of the unknowns in $\mathbb{R}^{NA}$ has been done in such a way that every $v_T \in T$ has only the first NT components which are, *a priori*, different from zero, while every $v_k \in K$ has only the last NK components (*a priori*) different from zero. With a (quite natural) abuse of notation we shall therefore, *when convenient*, treat v_T as an element of $\mathbb{R}^{NT}$ (discarding the last NK components which are identically zero). Similarly we shall treat, when convenient, v_K as an element of $\mathbb{R}^{NK}$ (discarding the first NT components) so that, for $v = v_T + v_K$ we can write:

$$Av = (A_{TT}v_T + A_{TK}v_K) + (A_{KT}v_T + A_{KK}v_K) \tag{3.7}$$

In (3.7) the first term of the right-hand side belongs to T and the second one belongs to K. Similarly, the matrix B will have the form

$$B = (B_T \quad B_K) \tag{3.8}$$

with

$$Bv = B_T v_T + B_K v_K, \tag{3.9}$$

always with the notation (3.5). Note now that from (3.3) and the definition of T we will have

$$Bv_K = B_K v_K = 0 \qquad \forall v_K \in K \tag{3.10}$$

and

$$Bv_T = B_T v_T = 0 \qquad \text{iff} \quad v_T = 0, \tag{3.11}$$

so that (3.9) can actually be written as

$$Bv = B_T v_T \tag{3.12}$$

We also have

$$B^t q = B_T^t q \in T, \quad \forall q \in \mathbb{R}^{NB}. \tag{3.13}$$

With a similar splitting for the right-hand side $f = f_T + f_K$ the original system (3.2) can now be written as

$$\left.\begin{array}{lcl} A_{TT}u_T + A_{TK}u_K + B_T^t p & = & f_T \\ A_{KT}u_T + A_{KK}u_K & = & f_K \\ B_T u_T & = & g \end{array}\right\} \tag{3.14}$$

The conditions for the solvability of (3.14) (and hence of (3.2)) are now clear: we need that (i) the equation $B_T u_T = g$ is solvable for every $g \in \mathbb{R}^{NB}$, (ii) the equation $A_{KK}u_K = f_K$ is solvable for every $f_K \in K$ and (iii) the equation $B_T^t p = f_T$ is solvable for every f_T in T. Condition (i) is equivalent to have that

$$B_T, \text{ as a mapping} : T \rightarrow \mathbb{R}^{NB}, \text{ is invertible.} \tag{3.15}$$

On the other hand, condition (ii) is equivalent to

$$A_{KK}, \text{ as a mapping} : K \to K, \text{ is invertible.} \tag{3.16}$$

Note now that, from the definition of B_T and in particular from (3.11) we have that B_T is always injective, so that (3.15) implies $NT = NB$. Now we conclude that the matrix B_T, as a mapping: $T \to \mathbb{R}^{NB}$ is a non-singular square matrix, and therefore its transposed matrix B_T^t is also non-singular and (iii) is automatically satisfied.

We want now to express (3.15) and (3.16) in terms of the matrices A and B, and of the kernel K (defined in (3.3)). Condition (3.15) is clearly equivalent to

$$B^t p = 0 \Rightarrow p = 0, \tag{3.17}$$

while (3.16) can be written as

$$(u \in K \text{ and } v^t A u = 0 \qquad \forall v \in K) \Rightarrow u = 0. \tag{3.18}$$

Conditions (3.17) and (3.18) are necessary and sufficient for the solvability of (3.2) for every right-hand side $f \in \mathbb{R}^{NA}$ and $g \in \mathbb{R}^{NB}$. We can summarize the above results in the following proposition.

PROPOSITION 3.1 Let A be an $NA \times NA$ square matrix and let B be an $NB \times NA$ matrix, and let K (the kernel of B) be defined as in (3.3). The linear system (3.2) is uniquely solvable for every $f \in \mathbb{R}^{NA}$ and for every $g \in \mathbb{R}^{NB}$ if and only if conditions (3.17) and (3.18) are satisfied. □

Note that, in particular, condition (3.17) implies

$$NA = NK + NT = NK + NB \geq NB \tag{3.19}$$

which is (obviously) a *necessary* condition for the solvability of (3.2). However, we now recognize that the use of (3.19) as a test for solvability (or, worse, for stability) is too simplistic and hence *misleading.* Note also that, if A is symmetric and positive semi-definite, then (3.18) can be expressed by the easier form

$$v^t A v > 0 \qquad \forall v \in K. \tag{3.20}$$

We address now the problem of *stability* of (3.2). In agreement with the approach of the previous section we might decide now from the very beginning to use dual norms for measuring the right-hand sides. Hence we assume that we have chosen a norm $\| \ \|_V$ in $\mathbb{R}^{NA}$ and a norm $\| \ \|_Q$ in $\mathbb{R}^{NB}$ and define the stability constant S as the smallest constant such that

$$\frac{\| \delta u \|_V + \| \delta p \|_Q}{\| u \|_V + \| p \|_Q} \leq S \frac{\| \delta f \|_{DV} + \| \delta g \|_{DQ}}{\| f \|_{DV} + \| g \|_{DQ}} \tag{3.21}$$

for all $u, p, \delta u, \delta p$, and $f, g, \delta f, \delta g$ with $Au + B^t p = f; Bu = g; A\delta u + B^t \delta p = \delta f$ and $B\delta u = \delta g$. From the previous section we have again $S \geq 1$.

REMARK 3.1 Definition (3.21) coincides with (2.11) if we take

$$\| (u,p) \|_L = \| u \|_V + \| p \|_Q \tag{3.22}$$

$$\| (f,g) \|_R = \| f \|_{DV} + \| g \|_{DQ} \, . \tag{3.23}$$

We notice that in this case the norm $\| \ \|_R$ is not the dual norm of $\| \ \|_L$. Actually we have

$$\| (f,g) \|_{DL} = \max(\| f \|_{DV}, \| g \|_{DQ}). \tag{3.24}$$

However one can easily check that

$$\| (f,g) \|_{DL} \leq \| (f,g) \|_{R} \leq 2 \| (f,g) \|_{DL} \tag{3.25}$$

so that the conditions for the uniform stability are still as discussed in the previous section. □

Our aim is now to give conditions on a sequence of problems (3.2) in order to have S uniformly bounded. We might of course use, for instance, (2.34) and (2.35) as in the previous section (since we are dealing here with a particular case of the previous discussion). However, we prefer to have separate conditions on the (sequence of) matrices A and B, as we did for the solvability problem. This, actually, is much more convenient in actual applications.

We assume, for the sake of simplicity, that there exist two constants k_A and k_B such that

$$v^t A u \leq k_A \| v \|_V \ \| u \|_V \qquad \forall v, u \tag{3.26}$$

$$v^t B^t q \leq k_B \| v \|_V \ \| q \|_Q \qquad \forall v, q \tag{3.27}$$

with k_A and k_B uniformly bounded from above and from below. In actual applications, (3.26) and (3.27), are easily fulfilled with the "natural choice" for the norms $\| \ \|_V$ and $\| \ \|_Q$. Notice that (3.26) and (3.27) immediately imply that A has norm $\leq k_A$ from $\| \ \|_V$ into $\| \ \|_{DV}$ (as in the previous section, Proposition 2.1). Similarly B_T has norm $\leq k_B$ from $\| \ \|_V$ into $\| \ \|_{DQ}$. On the other hand (3.26) and (3.27) also imply

$$\| M \|_{LR} \leq k_A + 2k_B \tag{3.28}$$

for M given in (3.1) and the norms $\| \quad \|_L$ and $\| \quad \|_R$ as in (3.22), (3.23). Hence, in view of (2.17) we have only to control $\| M^{-1} \|_{RL}$. Assuming that M is invertible (and hence, by (3.14) A_{KK} and B_T are also invertible) we have easily from (3.14) that

$$\| u_T \|_V \leq \| B_T^{-1} \| \; \| g \|_{DQ} \tag{3.29}$$

$$\begin{aligned} \| u_K \|_V &\leq \| A_{KK}^{-1} \| (\| f_K \|_{DV} + \| A_{KT} u_T \|_{DV}) \\ &\leq \| A_{KK}^{-1} \| (\| f_K \|_{DV} + k_A \| u_T \|_V) \end{aligned} \tag{3.30}$$

$$\begin{aligned} \| p \|_Q &\leq \| (B_T^t)^{-1} \| (\| f_T \|_{DV} + \| A_{TT} u_T \|_{DV} + \| A_{TK} u_K \|_{DV}) \leq \\ &\leq \| (B_T^t)^{-1} \| (\| f_T \|_{DV} + k_A(\| u_T \|_V + \| u_K \|_V)) \end{aligned} \tag{3.31}$$

where

$$\| B_T^{-1} \| = \sup_g \frac{\| B_T^{-1} g \|_V}{\| g \|_{DQ}} \tag{3.32}$$

$$\| A_{KK}^{-1} \| = \sup_{v \in K} \frac{\| A_{KK}^{-1} v \|_V}{\| v \|_{DV}} \tag{3.33}$$

and

$$\| (B_T^t)^{-1} \| = \sup_{f_T} \frac{\| (B_T^t)^{-1} f_T \|_Q}{\| f_T \|_{DV}}. \tag{3.34}$$

Substituting (3.29) into (3.30) and then (3.29) and (3.30) into (3.31) one obtains

$$\| (u,p) \| \leq C(\| B_T^{-1} \|, \| A_{KK}^{-1} \|, \| (B_T^t)^{-1} \|, k_A) \| (f,g) \| \tag{3.35}$$

Since, as it is easy to check,

$$\| B_T^{-1} \| = \| (B_T^t)^{-1} \| \tag{3.36}$$

it follows from (3.35) that, in order to have a uniform bound for S (in (3.21)), we only need that $\| B_T^{-1} \|$ (or $\| (B_T^t)^{-1} \|$) and $\| A_{KK}^{-1} \|$ are uniformly bounded from above. In order to express this condition in terms of the matrices A, B (and of the kernel K as defined in (3.3)) we shall rather write that $\| (B_T^t)^{-1} \|^{-1}$ and $\| A_{KK}^{-1} \|^{-1}$ are uniformly bounded from below by some positive constant. Actually we have

$$\begin{aligned}
\| (B_T^t)^{-1} \|^{-1} &= (\text{use } (3.35)) = (\sup_{f_T} \frac{\| (B_T^t)^{-1} f_T \|_Q}{\| f_T \|_{DV}})^{-1} = \\
&= \inf_{f_T} \frac{\| f_T \|_{DV}}{\| (B_T^t)^{-1} f_T \|_Q} = \\
&= (\text{use } f_T = B_T^t q) = \inf_q \frac{\| B_T^t q \|_{DV}}{\| q \|_Q} = \\
&= (\text{use } (3.12)) = \inf_q \frac{\| B^t q \|_{DV}}{\| q \|_Q} = \\
&= (\text{use } (2.25)) = \inf_q \sup_v \frac{v^t B^t q}{\| q \|_Q \ \| v \|_V} = \\
&= \inf_q \sup_v \frac{q^t B v}{\| v \|_V \ \| q \|_Q}
\end{aligned} \tag{3.37}$$

and

$$\begin{aligned}
\| A_{KK}^{-1} \|^{-1} &= (\text{use } (3.33)) = (\sup_{v \in K} \frac{\| A_{KK}^{-1} v \|_V}{\| v \|_{DV}})^{-1} = \\
&= \inf_{v \in K} \frac{\| v \|_{DV}}{\| A_{KK}^{-1} v \|_V} = \\
&= (\text{use } v = A_{KK} u) = \inf_{u \in K} \frac{\| A_{KK} u \|_{DV}}{\| u \|_V} = \\
&= (\text{use } (2.25)) = \inf_{u \in K} \sup_{z \in K} \frac{z^t A_{KK} u}{\| u \|_V \ \| z \|_V} = \\
&= (\text{use } (3.7)) = \inf_{u \in K} \sup_{z \in K} \frac{z^t A u}{\| u \|_V \ \| z \|_V}.
\end{aligned} \tag{3.38}$$

From (3.37)), (3.38) and the previous discussion we have now the following proposition.

PROPOSITION 3.2. Assume that we are given a sequence of problems of type (3.2). Assume that the matrices A and B satisfy (3.26) and (3.27) with k_A and k_B uniformly bounded. The stability constant S in (3.21) will then be unif [illegible] nded if and only if there exist two positive constants α and β such that

$$\inf_{u\in K}\sup_{v\in K}\frac{v^tAu}{\| v \|_V \ \| u \|_V} \ge \alpha > 0 \tag{3.39}$$

and

$$\inf_{q}\sup_{v}\frac{q^tBv}{\| v \|_V \ \| q \|_Q} \ge \beta > 0. \tag{3.40}$$

for every problem of the sequence.

REMARK 3.2 If every matrix A is symmetric and positive semi-definite, then (3.39) takes the simpler form

$$\exists\alpha > 0 \text{ such that } v^tAv \ge \alpha \| v \|_V^2 \qquad \forall v \in K \tag{3.41}$$

with K (as in (3.39)) always given by (3.3). In some applications (typically in the solution of Stokes fluid flow problems) the matrices A will be positive definite and satisfy (3.41) for all v in $I\!R^{NA}$. This led some authors to consider (3.40) as *the* condition for stability and convergence of mixed methods, which obviously is not the case. For instance, in the analysis of the mixed (σ, u) formulation of elasticity problems in the nearly incompressible case, condition (3.39) is more delicate to enforce than (3.40). On the same erroneous trend, some authors seem incapable of distinguishing between (2.35) (which is a condition on the whole matrix M) and (3.40) (which is a condition on the rectangular submatrix B of a special case of the matrix M, namely (3.1)) □

Let us consider now, as we did in the previous section, an abstract continuous problem and its Galerkin approximation. Assume that we are given two Hilbert spaces V and Q and two bilinear forms $a(u,v)$ (on $V \times V$) and $b(v,p)$ (on $V \times Q$). We assume from the beginning that the two forms are *continuous* in the sense that there exist two positive constants k_a and k_b such that

$$a(u,v) \leq k_a \parallel u \parallel_V \ \parallel v \parallel_V \qquad \forall u, v \in V \tag{3.42}$$

and

$$b(v,p) \leq k_b \parallel v \parallel_V \ \parallel p \parallel_Q \qquad \forall v \in V, p \in Q. \tag{3.43}$$

We can also introduce a kernel

$$K = \{v \in V \text{ such that } b(v,q) = 0 \qquad \forall q \in Q\} \tag{3.44}$$

which is the continuous version of the kernel K defined by (3.3). For the sake of simplicity we shall also assume that $a(u,v)$ is symmetric and positive semi-definite, that is

$$a(u,v) = a(v,u) \qquad \forall u, v \in V \tag{3.45}$$

$$a(v,v) \geq 0 \qquad \forall v \in V. \tag{3.46}$$

Finally, in analogy with (3.40) and (3.41) we make the following assumptions:

$$\exists\, \bar{\alpha} > 0 \text{ such that } a(v,v) \geq \bar{\alpha} \parallel v \parallel_V^2 \quad \forall\, v \in K \tag{3.47}$$

$$\exists\, \bar{\beta} > 0 \text{ such that } \inf_{q\in Q} \sup_{v\in V} \frac{b(v,q)}{\| v \|_V \ \| q \|_Q} \geq \bar{\beta}. \tag{3.48}$$

We have the following existence and uniqueness theorem.

THEOREM 3.1 [5] Assume (3.42) - (3.48). For every $f \in V'$ and for every $g \in Q'$, where V' and Q' are the dual spaces of V and Q, respectively, there exists a unique pair (u,p) in $V \times Q$ such that

$$\left\{ \begin{array}{lcll} a(u,v) + b(v,p) & = & f(v) & \forall\, v \in V \\ b(u,q) & = & g(q) & \forall\, q \in Q \end{array} \right. \tag{3.49}$$

□

Assume now that we are given a sequence (V_h, Q_h) of finite dimensional subspaces of V and Q respectively, and consider the finite dimensional approximations of (3.49):

$$\left\{ \begin{array}{lll} \text{find } u_h \in V_h \text{ and } \ p_h \in Q_h \text{ such that} & & \\ a(u_h, v_h) + b(v_h, p_h) & = f(v_h) & \forall v_h \in V_h \\ b(u_h, q_h) & = g(q_h) & \forall q_h \in Q_h. \end{array} \right. \tag{3.50}$$

It will also be convenient to introduce the finite dimensional kernels

$$K_h = \{v_h \in V_h, \text{ such that } b(v_h, q_h) = 0\, \forall\, q_h \in Q_h\}. \tag{3.51}$$

It is clear that, by choosing bases in V_h and Q_h, (3.50) can be written in the form (3.2). As a consequence, the solvability conditions for (3.50) will be:

$$a(v_h, v_h) > 0 \quad \forall v_h \in K_h \tag{3.52}$$

$$\{b(v_h, q_h) = 0 \quad \forall v_h \in V_h\} \Rightarrow q_h = 0 \tag{3.53}$$

as it can easily be deduced from (3.20) and (3.17). The uniform stability conditions become now

$$\exists\, \alpha^* > 0 \text{ such that } a(v_h, v_h) \geq \alpha^* \| v_h \|_V^2 \ \forall\, v_h \in K_h \tag{3.54}$$

$$\exists\, \beta^* > 0 \text{ such that } \inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\| v_h \|_V \ \| q_h \|_Q} \geq \beta^* \tag{3.55}$$

with α^* and β^* independent of h. It is clear that (3.54) and (3.55) are just a different way of writing (3.41) and (3.40). It is also clear that (3.54) implies (3.52), and (3.55) implies (3.53), so that stability implies solvability.

As far as error estimates are concerned we have the following theorem.

THEOREM 3.2 [5] Assume that the sequence of subspaces (V_h, Q_h) satisfies (3.54) and (3.55). Then problem (3.50) has a unique solution (u_h, p_h) for every $h > 0$. Moreover there exists a constant $c > 0$, depending only on k_a (3.42), k_b (3.43), α^* (3.54) and β^* (3.55) such that:

$$\| u - u_h \|_V + \| p - p_h \|_Q \ \leq \ C\{\inf_{v_h \in V_h} \| u - v_h \|_V + \inf_{q_h \in Q_h} \| p - q_h \|_Q\} \tag{3.56}$$

where (u, p) is the solution of (3.49).

PROOF. We shall only sketch the proof, which is based on a classical "stability-consistency" argument. Let u_h^* and p_h^* be the best approximation one can have for u and p (respectively) in the subspaces, that is

$$\| u - u_h^* \|_V = \inf_{v_h \in V_h} \| u - v_h \|_V \tag{3.57}$$

$$\| p - p_h^* \|_Q = \inf_{q_h \in Q_h} \| p - q_h \|_Q \tag{3.58}$$

Let now

$$\tilde{f}(v_h) := a(u_h^*, v_h) + b(v_h, p_h^*) \tag{3.59}$$

$$\tilde{g}(q_h) := b(u_h^*, q_h) \tag{3.60}$$

and notice that

$$(f - \tilde{f})(v_h) = a(u - u_h^*, v_h) + b(v_h, p - p_h^*), \tag{3.61}$$

$$(g - \tilde{g})(q_h) = b(u - u_h^*, q_h). \tag{3.62}$$

Notice finally that $(u_h - u_h^*, p_h - p_h^*)$ solves a problem of type (3.50) with right-hand side given by $(f - \tilde{f}, g - \tilde{g})$. The stability of (3.50) implies that

$$\begin{aligned} \| u_h - u_h^* \|_V &+ \| p_h - p_h^* \|_Q \leq C_1(\| f - \tilde{f} \|_{DV} + \| g - \tilde{g} \|_{DQ}) = \\ &= C_1\{\sup_{v_h} \frac{(f - \tilde{f})(v_h)}{\| v_h \|_V} + \sup_{q_h} \frac{(g - \tilde{g})(q_h)}{\| q_h \|_Q}\} \leq \\ &\leq C_2\{\| u - u_h^* \|_V + \| p - p_h^* \|_Q\} \end{aligned} \tag{3.63}$$

with C_1 and C_2 depending only on $\alpha^*, \beta^*, k_a, k_b$. From (3.62) and the triangle inequality we have now

$$\| u - u_h \|_V + \| p - p_h \|_Q \leq (1 + C_2)(\| u - u_h^* \|_V + \| p - p_h^* \|_Q) \tag{3.64}$$

and (3.64) with (3.56) and (3.57) gives (3.55). □

We end this section with some observations regarding penalty methods applied to systems of the form (3.2). For the sake of simplicity, assume that $NA = 3, NB = 2$ and that M has the form

$$M = \begin{pmatrix} \alpha_1 & 0 & 0 & \beta_1 & 0 \\ 0 & \alpha_2 & 0 & 0 & \beta_2 \\ 0 & 0 & \alpha_3 & 0 & 0 \\ \beta_1 & 0 & 0 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 & 0 \end{pmatrix} \tag{3.65}$$

For a more realistic situation we have to think of (3.65) as a block partitioning of M. It is clear that the system (3.2) splits now into

$$\left\{ \begin{array}{lcl} \alpha_i u_i + \beta_i p_i & = & f_i \\ \beta_i u_i & = & g_i \end{array} \right. (i = 1,2) \tag{3.66}$$

and

$$\alpha_3 u_3 = f_3 \tag{3.67}$$

If one of the β_i vanishes then (3.17) is violated and M is singular. If instead $\beta_i \neq 0$ $(i = 1,2)$ then

$$K = \{0, 0, u_3), \quad u_3 \in I\!R\} \tag{3.68}$$

and $\alpha_3 \neq 0$ satisfies (3.18). Assuming that all the α_i $(i = 1,2,3)$ are bounded away from zero (for the sake of simplicity) we have only to consider systems of type (3.66) that we consider through their typical representative:

$$\left\{ \begin{array}{lcl} \alpha u + \beta p & = & f \\ \beta u & = & g \end{array} \right. \tag{3.69}$$

A penalty approach to (3.69) consists in finding, for $\epsilon > 0$ (and "small") the solution of

$$\begin{cases} \alpha u_\epsilon + \beta p_\epsilon & = & f \\ \beta u_\epsilon - \epsilon p_\epsilon & = & g \end{cases} \tag{3.70}$$

which is given by

$$u_\epsilon = \frac{\epsilon f + \beta g}{\epsilon\alpha + \beta^2}\ ;\ p_\epsilon = \frac{\beta f - \alpha g}{\epsilon\alpha + \beta^2}. \tag{3.71}$$

It is clear that (for $\alpha \neq 0$) the solution (3.70) always exists, even for $\beta = 0$. However, for $\beta = 0, p_\epsilon \to \infty$ as $\epsilon \to 0$ when $g \neq 0$. On the other hand, in some applications (as, for instance, incompressibility conditions with zero Dirichlet boundary conditions), we have $g = 0$, and the situation improves. For $g = 0$ (3.71) becomes

$$u_\epsilon = \frac{\epsilon f}{\epsilon\alpha + \beta^2}\ ;\ p_\epsilon = \frac{\beta f}{\epsilon\alpha + \beta^2} \tag{3.72}$$

For $\beta = 0$ (3.72) gives

$$u_\epsilon = \frac{f}{\alpha};\quad p_\epsilon = 0 \tag{3.73}$$

which is a nice result. Actually a closer look at (3.69) for $\beta = g = 0$ shows a singular but compatible system with solution $u = f/\alpha, p =$ undetermined. Clearly (3.73) gives the solution of minimum norm.

Let us now consider the case $g = 0$ and β very small. The system (3.69) will have a very large stability constant. However, if we look only at the u_ϵ component of (3.72) we have

$$u_\epsilon \to 0 \text{ for } \epsilon \to 0 \text{ } (\beta \text{ fixed}) \tag{3.74}$$

and u_ϵ is uniformly bounded (in ϵ) as ϵ goes to zero. On the other hand

$$u_\epsilon \rightarrow \frac{f}{\alpha} \text{ for } \beta \rightarrow 0 \ (\epsilon \text{ fixed}) \tag{3.75}$$

which shows that we have difficulties to interpret the results even if u_ϵ is computed as a number of reasonable size. A look at the p_ϵ part of the solution shows that

$$p_\epsilon \rightarrow \frac{f}{\beta} \text{ for } \epsilon \rightarrow 0 \ (\beta \text{ fixed}) \tag{3.76}$$

If β is very small p_ϵ will be very large and this indicates that a change in discretization may be required.

In a practical analysis, there will generally be only a limited number of the β_i's that are small. Hence, only the corresponding p_i components will be large, and this can explain the appearance of the so-called checker-board modes that appear, even when u_ϵ behaves nicely. Note also that, when solving with the penalty approach (for $g = 0$) a small β_i can be more dangerous than a $\beta_i = 0$, as shown by (3.73) compared with (3.74) and (3.76).

4. EXAMPLES OF APPLICATIONS

ᵀn this section we present two examples that demonstrate the theory we have presented in the earlier part of the paper, and we also present a numerical procedure to test whether the inf-sup condition is satisfied for a finite element formulation to solve Stokes fluid flow problems.

4.1 Mixed Methods for Linear Second-Order Elliptic Problems

We start here with a very simple example to show the importance of the condition (3.54) ($\mathcal{K}_h$-ellipticity). Consider the mixed formulation of the model problem

$$\begin{cases} \psi'' = 1 & = \quad \text{in }]-1,1[\\ \psi(-1) = \psi(1) & = 0 \end{cases} \tag{4.1}$$

The solution is clearly

$$\psi(x) = \frac{x^2-1}{2}. \tag{4.2}$$

Introducing the additional variable

$$\sigma = \psi' \tag{4.3}$$

the mixed formulation of (4.1) reads now

$$\int_{-1}^{1} \sigma\tau \;\; dx + \int_{-1}^{1} \psi\tau' dx = 0 \quad \forall\tau \tag{4.4}$$

$$\int_{-1}^{1} \sigma' \, \phi \, dx = \int_{-1}^{1} \phi \;\; dx \quad \forall\phi \tag{4.5}$$

which is clearly of the form (3.49) with

$$V = \{\tau \in L^2(]-1,1[),\ \ \tau' \in L^2(]-1,1[\},\ \ \| \tau \|_V^2 = \| \tau \|_0^2 + \| \tau' \|_0^2 \tag{4.6}$$

$$Q = L^2(]-1,1[),\ \ \| \phi \|_Q = \| \phi \|_0 \tag{4.7}$$

$$a(\sigma,\tau) = \int_{-1}^{1} \sigma\, \tau\, dx; \quad b(\tau,\psi) = \int_{-1}^{1} \psi \tau' dx \tag{4.8}$$

where in (4.6) and (4.7) we used

$$\| v \|_0^2 := \int_{-1}^{1} v^2(x) dx. \tag{4.9}$$

Note how the form of a and b in (4.8) easily determines the norms (4.6) and (4.7) which are needed to have (3.42) and (3.43).

Let us check, as an exercise, that our problem satisfies (3.47) and (3.48). We have first to find what is K, as defined by (3.44). We have

$$\{\int_{-1}^{1} \phi \tau' dx = 0\ \forall\ \phi\} \Longleftrightarrow \tau' = 0 \Longleftrightarrow \tau = \text{ constant} \tag{4.10}$$

so that K contains only the constant functi . For $\tau \epsilon K$ we have

$$a(\tau,\tau) = \| \tau \|_0^2 = \| \tau \|_V^2 \quad (\text{since } \tau' = 0) \tag{4.11}$$

and therefore (3.47) holds with $\bar{\alpha} = 1$.

Let us now turn to (3.48). For every $\bar{\phi} \in L^2(]-1,1[)$ we can set

$$\bar{\tau}(x) = \int_0^x \bar{\phi}(t) dt \tag{4.12}$$

We then obviously have

$$b(\bar{\tau}, \bar{\phi}) = \int_{-1}^{1} \bar{\phi}^2 = \| \bar{\phi} \|_0^2 \tag{4.13}$$

$$\| \bar{\tau}' \|_0^2 = \| \bar{\phi} \|_0^2 \tag{4.14}$$

Furthermore

$$\| \bar{\tau} \|_0^2 \leq \| \bar{\phi} \|_0^2 \, . \tag{4.15}$$

and hence we have

$$\sup_{\tau} \frac{b(\tau, \bar{\phi})}{\| \tau \|_V} \geq \frac{b(\bar{\tau}, \bar{\phi})}{\| \bar{\tau} \|_V} = \frac{\| \bar{\phi} \|_0^2}{(\| \bar{\tau} \|_0^2 + \| \bar{\tau}' \|_0^2)^{1/2}} \geq$$

$$\geq (\text{ use (4.14) and (4.15)}) \geq \frac{\| \bar{\phi} \|_0^2}{(\| \bar{\phi} \|_0^2 + \| \bar{\phi} \|_0^2)^{1/2}} = \frac{1}{\sqrt{2}} \| \bar{\phi} \|_0 \tag{4.16}$$

Since (4.16) holds for every $\bar{\phi}$ we obtain (3.48) with $\bar{\beta} = \frac{1}{\sqrt{2}}$.

Let us now consider the discretization of (4.4), (4.5). We take a decomposition of $]-1,\ 1[$ into N equal intervals and set

$$Q_h = \{\text{piecewise constants } (= \mathcal{L}_0^0 \text{ with the notation of } [1])\} \tag{4.17}$$

It would now be reasonable to take

$$V_h = \{\text{piecewise linear continuous functions } (= \mathcal{L}_1^1)\} \tag{4.18}$$

In this case it is easy to check that K_h is also reduced to the constant functions, and therefore (3.54) holds with $\alpha^* = 1$ (always by (4.11)). On the other hand the construction (4.12) still works, since $\bar{\phi} \in \mathcal{L}_0^0$ implies $\bar{\tau} \in \mathcal{L}_1^1$. Hence (3.55) also holds, with $\beta^* = 1/\sqrt{2}$, and the assumptions of Theorem 3.2 are fulfilled. In our particular case ($g \equiv 1$) it is also easy to prove that, for every decomposition, we have

$$\sigma_h(x) = x \qquad \text{in }]-1,1[\tag{4.19}$$

which is the exact solution.

Assume now, for our discussion, that we take a *larger* V_h, namely:

$$V_h = \{\text{piecewise quadratic continuous functions } (= \mathcal{L}_2^1)\}. \tag{4.20}$$

Setting

$$B_2 = \{\text{piecewise quadratic functions vanishing at the subdivision nodes}\} \tag{4.21}$$

we easily have

$$V_h = \mathcal{L}_2^1 = \mathcal{L}_1^1 \oplus B_2. \tag{4.22}$$

We can now make an observation which is of *general* validity: the choice of a larger V_h (with the same Q_h) makes the inf-sup condition (3.55) *easier* to satisfy and the $\mathcal{K}_h$-ellipticity condition (3.54) *more difficult* to satisfy (unless, obviously, the bilinear form a is $V-$elliptic: in such a case (3.54) is always satisfied for all choices of V_h and Q_h). In our case

$$\mathcal{K}_h = \{\tau \in \mathcal{L}_2^1, \text{ such that } \int_{-1}^{1} \phi\tau' dx = 0 \quad \forall \phi \in \mathcal{L}_0^0\} = \mathcal{K} \oplus B_2 \tag{4.23}$$

where $\mathcal{K}$ is the space of global constants as in (4.10). Let now, for every subinterval $I_k (k = 1, \cdots N)$, b_k be the second order polynomial vanishing at the endpoints of I_k and normalized in such a way that

$$\int_{I_k} b_k^2(x)dx = 1 \tag{4.24}$$

A simple computation shows that

$$\| b_k' \|_0^2 = 10/h \tag{4.25}$$

so that

$$a(b_k, b_k) = \| b_k \|_0^2 = 1 \tag{4.26}$$

and

$$\| b_k \|_V^2 = \| b_k \|_0^2 + \| b_k' \|_0^2 = 1 + 10/h \tag{4.27}$$

and condition (3.54) can only hold for

$$\alpha^* = h/(h+10) \tag{4.28}$$

which is not bounded uniformly from below. On the other hand, it is obvious that for every $\phi \in \mathcal{L}_0^0$ we have

$$\sup_{\tau \in \mathcal{L}_2^1} \frac{b(\tau, \phi)}{\| \tau \|_V} \geq \sup_{\tau \in \mathcal{L}_1^1} \frac{b(\tau, \phi)}{\| \tau \|_V} \geq \frac{1}{\sqrt{2}} \| \phi \|_Q \tag{4.29}$$

and (3.55), as predicted, is easier satisfied with a larger V_h. We are therefore facing a case where the inf-sup condition (3.55) is easily satisfied, but the K_h−ellipticity condition (3.54) holds only with $\alpha^* \sim h$. Notice that (3.52) is still satisfied so that the discrete problem will be uniquely solvable. Notice as well that we started with an effective discretization ($V_h = \mathcal{L}_1^1, Q_h = \mathcal{L}_0^0$), that gave the exact value for σ_h, and that we *enlarged* V_h (which, as we have seen, does not affect the ability to satisfy the inf-sup condition). The key question must now of course be: "how is our solution accuracy affected by the enlargement by V_h?"

The solution of the discrete problem: find $\sigma_h \in \mathcal{L}_2^1$ and $\psi_h \in \mathcal{L}_0^0$ such that

$$\int_{-1}^{1} \sigma_h \tau dx + \int_{-1}^{1} \psi_h \tau' dx = 0 \qquad \forall \tau \in \mathcal{L}_2^1, \tag{4.30}$$

$$\int_{-1}^{1} \phi \sigma_h' dx - \int_{-1}^{1} \phi dx = 0 \qquad \forall \phi \in \mathcal{L}_0^0, \tag{4.31}$$

can again be computed by hand. Namely, from (4.31) we obtain

$$\sigma_h(x) = x + c + \sum_{k=1}^{N} c_k b_k(x)$$

with c and c_k to be determined. Now $c = 0$ for symmetry reasons (the solution is unique) and choosing $\tau = b_k(x)$ in (4.30) yields

$$c_k = -\int_{I_k} x b_k(x) dx =: (x, b_k) \tag{4.32}$$

so that

$$\sigma_h(x) = x - \sum_{k=1}^{N} b_k(x)(x, b_k), \tag{4.33}$$

and the L^2 norm of the error $\sigma_h - \sigma = \sigma_h - x$ is given by

$$\| \sigma_h - \sigma \|_0^2 = \sum_{k=1}^{N} (x, b_k)^2 \tag{4.34}$$

which *does not tend to zero.* Hence our solution scheme with the enlarged V_h is not acceptable.

4.2 Analysis of Stationary Incompressible Fluids (the Stokes Problem)

We consider now the following model problem in a smooth domain $\Omega \subset \mathbb{R}^2$

$$\begin{aligned}&\text{find } u \in (H_0^1(\Omega))^2 \text{ and } p \in L^2(\Omega)/\mathbb{R} \text{ such that}\\ &\int_\Omega \text{grad } u : \text{grad } v d\Omega + \int_\Omega p \text{ div } v d\Omega = \int_\Omega f \cdot v \, d\Omega \quad \forall \ v \\ &\int_\Omega q \text{ div } u d\Omega = 0 \qquad \forall q\end{aligned} \tag{4.35}$$

which is again of type (3.49) for

$$V = (H_0^1(\Omega))^2 = \{v \in (L^2(\Omega))^2 \text{ such that grad } v \in (L^2(\Omega))^2 \text{ and } v|_{\partial\Omega} = 0\} \tag{4.36}$$

$$Q = L^2(\Omega)/\mathbb{R} = \{q \in L^2(\Omega), \int_\Omega q \, d\Omega = 0\} \tag{4.37}$$

and

$$a(u,v) = \int_\Omega \text{grad } u : \text{grad } v \, d\Omega; \;\; b(v,q) = \int_\Omega q \text{ div } v \, d\Omega. \tag{4.38}$$

Note again how the form of a and b in (4.38) easily determines the norms (4.36) and (4.37) which are needed for having (3.42) and (3.43).

Problem (4.35) is the variational formulation of the problem

$$\begin{cases} -\Delta u - \nabla p = f & \text{in } \ \Omega \\ \text{div } u = 0 & \text{in } \ \Omega \\ u = 0 & \text{on } \ \partial\Omega \end{cases} \tag{4.39}$$

which are the governing equations of an incompressible fluids. The well-known Poincaré inequality:

$$\exists c = c(\Omega) \quad \text{with} \quad \int_\Omega |v|^2 d\Omega \le c(\Omega) \int_\Omega |\text{grad } v|^2 d\Omega \ \ \forall \ \ v \in V \tag{4.40}$$

ensures now that

$$a(v,v) \ge (\frac{c}{c+1}) \parallel v \parallel_V^2 \qquad \forall v \in V \tag{4.41}$$

so that (3.47) holds with $\bar{\alpha} = c/(c+1)$ *for all* $v \in V$ and we do not need to be concerned with kernel $\mathcal{K}$. In particular, we have that (3.54) will also hold (with $\alpha^* = c/(c+1)$) for every choice of $V_h \subset V$ and $Q_h \subset Q$. Hence we can concentrate our attention on (3.48) and (3.55), that is on the inf-sup condition. As far as (3.48) is concerned we remark that we actually have

$$\exists \ \beta(\Omega) > 0 \text{ such that } \inf_{q \in Q} \ \sup_{v \in V} \frac{\int_\Omega q \text{ div } v \ d\Omega}{\parallel q \parallel_Q \ \parallel v \parallel_V} \ge \beta(\Omega) \tag{4.42}$$

which is a nontrivial result in functional analysis (see, e.g., [8,9]). We also notice that the following result obviously holds as an immediate consequence of (4.42): for every set $\mathcal{V}$ with

$$(H_0^1(\Omega))^2 \subseteq \mathcal{V} \subseteq (H^1(\Omega))^2 \tag{4.43}$$

we have

$$\inf_{q \in Q} \sup_{v \in \mathcal{V}} \frac{\int_\Omega q \text{ div } v \ d\Omega}{\parallel q \parallel_Q \ \parallel v \parallel_V} \ge \beta(\Omega) \tag{4.44}$$

since the supremum over $\mathcal{V}$ is obviously larger than the supremum over $V \equiv (H_0^1(\Omega))^2$. In a sense we can therefore say that the case of homogeneous Dirichlet boundary conditions is the most difficult to treat. This is the reason why we shall mainly concentrate on this case only.

Assume now that we are given a sequence of finite dimensional subspaces $V_h \in V$ and $Q_h \in Q$ and consider the discrete problem:

$$\begin{cases} \text{find } u_h \in V_h \text{ and } p_h \in Q_h \text{ such that :} \\ a(u_h, v_h) + b(v_h, p_h) = \int_\Omega f \cdot v_h d\Omega \quad \forall v_h \in V_h, \\ b(u_h, q_h) = 0 \qquad \forall q_h \in Q_h. \end{cases} \tag{4.45}$$

As we have observed above, we only have to check condition (3.55) in order to have solvability, stability and optimal error bounds. The following theorem, known as "Fortin's trick" (see ref. [10]) is often useful in order to prove (3.55)[11].

THEOREM 4.1 Assume that (3.48) holds, and assume that, for every h, we can build a linear operator $\Pi_h : V \rightarrow V_h$ with the following properties

$$b(v - \Pi_h v, q_h) = 0 \qquad \forall v \in V, \ \forall q_h \in Q_h \tag{4.46}$$

$$\exists \gamma > 0 \text{ such that } \| \Pi_h v \|_V \leq \gamma \| v \|_V \quad \forall v \in V \tag{4.47}$$

where γ is independent of h. Then (3.55) holds with $\beta^* = \bar{\beta}/\gamma$.

PROOF. We have for every $q_h \in Q_h$:

$$\begin{aligned} \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\| v_h \|_V} &\geq \sup_{v \in V} \frac{b(\Pi_h v, q_h)}{\| \Pi_h v \|_V} = \\ &= (\text{use } (4.46)) = \sup_{v \in V} \frac{b(v, q_h)}{\| \Pi_h v \|_V} \geq \\ &\geq \ (\text{use } (4.47)) \geq \sup_{v \in V} \frac{b(v, q_h)}{\gamma \| v \|_V} \geq \\ &\geq \ (\text{use } (3.48)) \geq \bar{\beta}/\gamma \| q_h \|_Q . \end{aligned} \tag{4.48}$$

where the first inequality holds since the image $\Pi_h(V)$ is contained in V_h. □

In many cases, it will actually be sufficient to prove that, for every $q_h \in Q_h$ we have

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\| v_h \|_V} \geq \kappa \| q_h - \bar{q}_h \|_0 \tag{4.49}$$

where κ is independent of h and

$$\bar{q}_h = \{L^2 - \text{projection of} q_h \text{ onto the space } \mathcal{L}_0^0 \text{ of piecewise constants}\} \tag{4.50}$$

Here for instance we present two classes of discretizations for which (4.49) implies (3.55).

PROPOSITION 4.1 Assume that $Q_h \subset C^\circ(\Omega)$ (and piecewise polynomial), and assume that V_h is locally first order accurate, in the sense that, for every $v \in V$ there exists a $v^I \in V_h$ with

$$\| v - v^I \|_{L^2} \leq c_1 h \| v \|_V \tag{4.51}$$

$$\| v^I \|_V \leq c_2 \| v \|_V . \tag{4.52}$$

Assume finally that the decomposition is quasi-uniform, in the sense that the maximum diameter h is bounded by $c_3 h_{\min}$ ($h_{\min}$ = minimum diameter of the element). Then (4.49) implies (3.55).

PROOF. We note that for every $q_h \in Q_h$ there exists a $\bar{v} \in V$ such that

$$\frac{b(\bar{v}, q_h)}{\| \bar{v} \|_V} \geq \bar{\beta} \| q_h \|_Q \tag{4.53}$$

Hence

$$\begin{aligned}
&\sup_{v_h\in V_h}\frac{b(v_h,q_h)}{\| v_h \|_V} \geq \frac{b(\bar{v}^I,q_h)}{\| \bar{v}^I \|_V} = (\pm\bar{v}) = \\
&= \frac{b(\bar{v}^I-\bar{v},q_h)}{\| \bar{v}^I \|_V} + \frac{b(\bar{v},q_h)}{\| \bar{v}^I \|_V} \geq \text{ (use (4.52))} \\
&\geq \frac{b(\bar{v}^I-\bar{v},q_h)}{c_2 \| \bar{v} \|_V} + \frac{b(\bar{v},q_h)}{c_2 \| \bar{v} \|_v} \geq \text{ (use (4.53))} \\
&\geq \frac{b(v^I-\bar{v},q_h)}{c_2 \| \bar{v} \|_V} + \bar{\beta} \| q_h \|_Q = \text{ (definition of } b) \\
&= \frac{\int_\Omega q_h \operatorname{div}(\bar{v}^I-\bar{v})d\Omega}{c_2 \| \bar{v} \|_V} + \bar{\beta} \| q_h \|_Q = \text{ (integrate by parts)} \\
&= -\frac{\int_\Omega(\bar{v}-\bar{v}^I)\cdot \operatorname{grad} q_h d\Omega}{c_2 \| \bar{v} \|_V} + \bar{\beta} \| q_h \|_Q \geq \\
&\geq \bar{\beta} \| q_h \|_Q - \frac{\| \bar{v}-\bar{v}^I \|_0}{c_2 \| \bar{v} \|_V} \| \operatorname{grad}\ q_h \|_0 \geq \\
&\geq \text{(use (4.51))}\ \bar{\beta} \| q_h \|_Q - \tfrac{c_1}{c_2} \| \operatorname{grad}\ q_h \|_0\, h.
\end{aligned} \tag{4.54}$$

Now a simple scaling argument (using the quasi-uniformity of the mesh) shows that

$$h \| \operatorname{grad} q_h \|_0 \leq c_4 \| q_h - \bar{q}_h \|_0 \tag{4.55}$$

and from (4.54) and (4.55) we have

$$\sup_{v_h\in V_h}\frac{b(v_h,q_h)}{\| v_h \|_V} \geq \bar{\beta} \| q_h \|_0 - c_5 \| q_h - \bar{q}_h \|_0 \tag{4.56}$$

It is now clear that (4.49) and (4.56) imply

$$\left(1+\frac{\kappa}{c_5}\right)\left(\sup_{v_h\in V_h}\frac{b(v_h,q_h)}{\| v_h \|_V}\right) \geq \frac{\bar{\beta}\cdot\kappa}{c_5} \| q_h \|_0 \tag{4.57}$$

and (3.55) holds with $\beta^* = \dfrac{\bar{\beta}\kappa/c_5}{1+\kappa/c_5}$. □

REMARK 4.1 The quasi-uniformity assumption is actually not necessary. We used it only in order to simplify the argument. See [1,12] for the general case.

PROPOSITION 4.2 Assume that we know that, for all $\bar{q}_h \in \mathcal{L}_0^0$

$$\sup_{v_h \in V_h} \frac{b(v_h, \bar{q}_h)}{\| v_h \|_V} \geq \gamma_1 \| \bar{q}_h \|_0 . \tag{4.58}$$

Then (4.49) implies (3.55)

PROOF. We have, for every $q_h \in Q_h$

$$\begin{aligned} \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\| v_h \|_V} = (\pm \bar{q}_h) &= \sup_{v_h \in V_h} \{ \frac{b(v_h, q_h - \bar{q}_h)}{\| v_h \|_V} + \frac{b(v_h, \bar{q}_h)}{\| v_h \|_V} \} \\ &\geq \sup_{v_h \in V_h} \frac{b(v_h, \bar{q}_h)}{\| v_h \|_V} - \sup_{v_h \in V_h} \frac{b(v_h, q_h - \bar{q}_h)}{\| v_h \|_V} \\ &\geq \gamma_1 \| \bar{q}_h \|_0 - \| q_h - \bar{q}_h \|_0 . \end{aligned} \tag{4.59}$$

Now from (4.59) and (4.49) we deduce (as in (4.57)) that

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\| v_h \|_V} \geq \frac{\kappa \gamma_1}{1 + \kappa} \| \bar{q}_h \|_0 . \tag{4.60}$$

Finally from (4.49) and (4.60) the result follows, since $\bar{q}_h$ and $q_h - \bar{q}_h$ are orthogonal in $L^2(\Omega)$.

REMARK 4.2 As we can see, (4.49) implies the inf-sup condition in an impressive number of cases: using (4.49) basically all the approximations with continuous pressure can be considered, as well as all the choices of V_h that give a stable pair when used with a piecewise constant pressure. This second case for instance holds true, if for triangular elements, V_h contains the space of piecewise quadratic functions, and if for quadrilateral elements, V_h contains the reduced (8 nodes) biquadratic functions. If we succeed in giving an easy test for (4.49) then we can treat a very wide number of cases. Such a test will be a consequence of proposition 4.3 below.

REMARK 4.3 More generally the condition we need on V_h (in order to have (4.58)) is the following: we can use as degrees of freedom the values $v \cdot n$ (normal component of velocity) at midpoints of edges (respectively, faces in $\mathbb{R}^3$). Indeed, if this is the case, we can (roughly) consider a "Fortin interpolator" Π_h such that

$$\int_e (v - \Pi_h v) \cdot n \, de = 0 \qquad \text{for all edges (faces) } e \tag{4.61}$$

and

$$\text{“}v = \Pi_h v\text{”} \quad \text{for the other degrees of freedom} \tag{4.62}$$

From (4.61) and (4.62) we have now, for every $\overline{q}_h \in \mathcal{L}_0^0$:

$$\int_K \begin{array}{l} \operatorname{div}(v - \Pi_h v)\overline{q}_h dK = (\text{ use Gauss theorem}) \; = \int_{\partial K} (v - \Pi_h v) \cdot n \overline{q}_h de = \\ (\text{use (4.61) on each } e) = 0, \end{array} \tag{4.63}$$

for every element K. This gives (4.46) for $Q_h = \mathcal{L}_0^0$, which is the essential step (through Theorem 4.1) in order to have (4.58). Note however that the actual proof of this fact has some more technicalities (see [1,13] for similar arguments). It is clear by now that, in designing a new element, to satisfy (4.49) is the essential step in almost every case.

PROPOSITION 4.3 Assume that P_1 and P_2 are unions of elements, with $P_1 \cap P_2 = \emptyset$. and that, for $i = 1, 2$, we have for all $q_h \in Q_h$

$$\sup_{\substack{v_h \in V_h \\ v_h = 0 \text{ in } \Omega \setminus P_i}} \frac{\int_{P_i} q_h \operatorname{div} v_h}{\| v_h \|_V} \geq \kappa_i \| q_h - \bar{q}_h \|_{L^2(P_i)} \tag{4.64}$$

then

$$\sup_{\substack{v_h \in V_h \\ v_h = 0 \text{ in } \Omega \setminus (P_1 \cup P_2)}} \frac{\int_{P_1 \cup P_2} q_h \operatorname{div} v_h}{\| v_h \|_V} \geq \kappa \| q_h - \bar{q}_h \|_{L^2(P_1 \cup P_2)} \tag{4.65}$$

for all $q_h \in Q_h$ and with

$$\kappa \geq \min(\kappa_1, \kappa_2). \tag{4.66}$$

PROOF. The proof is elementary. From (4.64) we have, for all $q_h \in Q_h$, two elements $v^i \in (H^1(\mathcal{P}_i))^2 \cap V_h (i = 1, 2)$, such that

$$b(v^i, q_h) = \kappa_i \parallel q_h - \bar{q}_h \parallel^2_{L^2(\mathcal{P}_i)} \tag{4.67}$$

and

$$\parallel v^i \parallel_V \leq \parallel q_h - \bar{q}_h \parallel_{L^2(\mathcal{P}_i)} \tag{4.68}$$

Taking $v = v^1 + v^2$ we have

$$b(v, q_h) = \sum_{i=1}^{2} \kappa_i \parallel q_h - \bar{q}_h \parallel^2_{L^2(\mathcal{P}_i)} \geq \text{(use (4.66))} \geq \kappa \parallel q_h - \bar{q}_h \parallel^2_{L^2(\mathcal{P}_1 \cup \mathcal{P}_2)} \tag{4.69}$$

and

$$\parallel v \parallel^2_V = \parallel v^1 \parallel^2_V + \parallel v^2 \parallel^2_V \leq \parallel q_h - \bar{q}_h \parallel^2_{L^2(\mathcal{P}_1 \cup \mathcal{P}_2)} \tag{4.70}$$

and (4.69) with (4.70) implies (4.65). □

REMARK 4.4 The condition $\mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset$, as we can see from the proof, is not crucial. Its only purpose is to avoid a factor 2 in (4.69) and (4.70). However, we always think of using the result in the "disjoint" case.

REMARK 4.5 In (4.64), (4.65) the choice of $\bar{q}_h$ as an element by element projection onto the space $\mathcal{L}^0_0$ of piecewise constants is unnecessary. We might as well use a "patch by patch" projection, that is we might assume that $\bar{q}_h$ is constant in every patch (and equal to the mean value of q_h). On the other hand, the choice $\bar{q}_h = 0$ (which would give directly the inf-sup condition without passing through Propositions 4.1 and 4.2) is not allowed. If q_h has zero mean value on $\mathcal{P}_1 \cup \mathcal{P}_2$ it does not necessarily have zero mean value separately on $\mathcal{P}_1$ and on $\mathcal{P}_2$. But (4.64) is unrealistic if the right-hand side does not have zero mean value. Finally let us note that, if $\bar{q}_h$ is the element by element projection and $\tilde{q}_h$ is the patch by patch projection then

$$\| q - \tilde{q}_h \|_{L^2} \geq \| q - \bar{q}_h \|_{L^2}$$

REMARK 4.6 Proposition 4.3 deals with two patches. It is clear that the argument applies as well to any finite number of patches: the smallest κ_i gives the global κ.

REMARK 4.7 It is very important to point out that conditions (4.64) do not depend on the size of the patches. Assume that we have, for a given patch, say, of size one

$$\sup_{v_h \in V_h(\mathcal{P})} \frac{\int_{\mathcal{P}} q_h \operatorname{div} v_h}{\| v_h \|_{H^1(\mathcal{P})}} \geq \| q_h - \tilde{q}_h \|_{L^2(\mathcal{P})} \tag{4.71}$$

for all $q_h \in Q_h(\mathcal{P})$, where $V_h(\mathcal{P})$ is a finite element subspace of $(H_0^1(\mathcal{P}))^2$ and $Q_h(\mathcal{P})$ is a finite element subspace of $L^2(\mathcal{P})$ and finally $\tilde{q}_h$ is the mean value of q_h on $\mathcal{P}$. If we shrink $\mathcal{P}$ to a small patch $\mathcal{P}^*$ of size h by the change of variable

$$\mathcal{P} \ni x = \xi / h, \xi \in \mathcal{P}^* \tag{4.72}$$

and if we change the finite element spaces accordingly, we have

$$\sup_{v_h^* \in V_h^*} \frac{\int_{\mathcal{P}^*} q_h^* \operatorname{div} v_h^*}{\| v_h^* \|_{H^1(\mathcal{P}^*)}} \geq \kappa \| q_h^* - \tilde{q}_h^* \|_{L^2(\mathcal{P}^*)} \tag{4.73}$$

exactly with the same κ as in (4.71).

The same is true if, instead of the change of variable (4.72) we apply any other change of variable which is "affine", that is with constant Jacobian: actually a distortion in the shape might change κ to some $\delta \cdot \kappa$, where δ depends on the amount of distortion (essentially if J is the Jacobian matrix, δ will depend on $|J^{-1}| \cdot \| J \|^2$ and $|J| \cdot \| J^{-1} \|^2$), and for reasonable distortions κ will remain greater than zero. □

Proposition 4.3 and all the remarks after it suggest now a strategy for checking the inf-sup condition, if we have a continuous pressure field or if we know already that the velocity space V_h under consideration can be used with piecewise constant pressures. Assume that we can find a finite number of "representative patches", $\mathcal{P}_1, \mathcal{P}_2, \cdots \mathcal{P}_R$ such that one can cover the decomposition with affine images of the $\mathcal{P}_i$'s. For each patch $\mathcal{P}_i$ we then check if the discrete problem is uniquely solvable on the patch (for both velocity and pressure) with homogeneous Dirichlet boundary conditions for velocities and obviously discarding the constant value of the pressure on the patch. If this is true, then a constant κ_i must exist: we do not need to compute it, we just want to know that it exists. Finally, the smallest κ_i will give the constant κ in (4.49) and the inf-sup condition will hold.

5. CONCLUDING REMARKS

Our objective in this paper was to review and discuss conditions for the stability of mixed finite element formulations. We also presented a numerical test that can be employed to check whether a given mixed finite element formulation for the general Stokes problem satisfies the mathematical conditions of stability and optimal error bounds.

While the general mathematical theory for mixed formulations is quite well established, both for the continuous and discretized problems, the actual detailed use of that theory for the design and analysis of mixed finite element formulations can be a very difficult task. We note that quite effective mixed finite elements that satisfy the mathematical conditions of stability and optimal error bounds are available for the solution of incompressible fluid flow [1] and the analysis of incompressible or almost incompressible solid media [1,14,15]. However, the situation is quite different, for example, in the field of analysis of plate and shell structures [16]. Here numerous mixed finite elements have been proposed but mathematical analyses are hardly available. Indeed the construction of effective mixed plate and shell elements that can be analysed and satisfy the mathematical conditions of stability and optimal error bounds is very difficult, and such elements are now under active research, see for example references [16-18].

REFERENCES

1. Brezzi, F. and Fcrtin, M., **Mixed and Hybrid Finite Element Methods**, (book to appear).

2. Bathe, K.J., **Finite Element Procedures in Engineering Analysis**, Prentice-Hall, 1982.

3. Zienkiewicz, O.C.; Qu, S.; Taylor, R.L. and Nakazawa, S., "The Patch Test for Mixed Formulations", *Int. J. Num. Meth. in Eng.*, **23**, 1873-1883, 1986.

4. Zienkiewicz, O.C. and Lefebvre, D., "Three-Field Mixed Approximation and the Plate Bending Problem", *Communications in Applied Numerical Methods*, **3**, 301-309, 1987.

5. Brezzi, F., "On the Existence, Uniqueness, and Approximation of Saddle-Point Problems Arising from Lagrange Multipliers", *R.A.I.R.O.*, **B-R2**, 129-151, 1974.

6. Babuška, I., "The Finite Element Method with Lagrangian Multipliers", *Numer. Math.*, **20**, 179-192, 1973.

7. Arnold, D.N., "Discretization by Finite Elements of a Model Parameter Dependent Problem", *Numer. Math.*, **37**, 405-421, 1981.

8. Ladyzhenskaya, O., **The Mathematical Theory of Viscous Incompressible Flows**, Gordon and Breach, 1969.

9. Temam, R., **Navier-Stokes Equations**, North Holland, 1977.

10. Fortin, M., "An Analysis of the Convergence of Mixed Finite Element Method", *R.A.I.R.O., Anal. Numér.*, **11**, 341-354, 1977.

11. Brezzi, F. and Bathe, K.J., "Studies of Finite Element Procedures – The Inf-Sup Condition, Equivalent Forms and Applications", in **Reliability of Methods for Engineering Analysis**, (K.J. Bathe and D.R.J. Owen, Eds.), Pineridge Press, 1986.

12. Verfürth, R., "Error Estimates for a Mixed Finite Element Approximation of the Stokes Equations", *R.A.I.R.O.*, **18**, 175-182, 1984.

13. Fortin, M., "Old and New Finite Elements for Incompressible Flows", *Int. J. Num. Meth. in Fluids*, **1**, 347-364, 1981.

14. Oden, J.T. and Kikuchi, N., "Finite Element Methods for Constrained Problems in Elasticity", *Int. J. Num. Meth. in Eng.*, **18**, 701-725, 1982.

15. Sussman, T. and Bathe, K.J., "A Finite Element Formulation for Nonlinear Incompressible Elastic and Inelastic Analysis", *J. Computers and Structures*, **26**, No. 1/2, 357-409, 1987.

16. Noor, A., et al. (Eds.), **Analytical and Computational Models of Shells**, ASME Special Publication, 1989.

17. Brezzi, F.; Bathe, K.J. and Fortin, M., "Mixed-Interpolated Elements for Reissner/Mindlin Plates", *Int. J. Num. Meth. in Eng.*, **28**, 1787-1801, 1989.

18. Bathe, K.J.; Cho, S.W.; Bucalem, M.L. and Brezzi, F., "On Our MITC Plate Bending/Shell Elements", in **Analytical and Computational Models of Shells**, (A.K. Noor, et. al., Eds.), ASME Special Publication, 1989.

A Posteriori Error Estimation

Richard E. Ewing

Departments of Mathematics, Petroleum Engineering

and Chemical Engineering

University of Wyoming

Abstract

I. INTRODUCTION

The emergence of new supercomputer architectures is helping to revolutionize the modeling process for large-scale physical problems. The advance in computational capabilities has allowed the incorporation of more physics in the model, thereby greatly increasing the complexity of the mathematical models and the simulation process. If decisions are to be made based on the results of a simulation, we must be able to determine the reliability of the modeling process. Both *a priori* and *a posteriori* error estimates are important in assessing the accuracy of the simulation and in helping to determine an adaptive strategy to improve the accuracy where needed. Unfortunately, none of the existing large-scale commercial codes have rigorous error estimation and adaptive improvement capabilities. The role of *a posteriori* error estimation in grid refinement and adaptivity is considered here.

There are four aspects of reliability which are tightly interrelated. First, an understanding of the accuracy of the modeling process must be obtained. The choice of the model, boundary conditions, and computational domain can greatly affect the properties of the solution and the effectiveness of the modeling process. Babuška has given an excellent introduction to this aspect of modeling in [1].

Once a model has been set, the properties of that model, including existence, uniqueness, and regularity, must be understood. Then, a choice of discretization scheme is made and *a priori* error estimates are obtained. *A priori* estimates, based upon knowledge of the general properties of solutions for the model equations and the approximation properties of the discretization methods, can give us a qualitative assessment of the error and the asymptotic rate of convergence as the number of degrees of freedom in the approximation tends to infinity. *A priori* estimates provide pessimistic indications of the error based upon upper bounds for Sobolev norms of the solution. However, they usually do not provide much information about the actual error in the discrete approximation. Nevertheless, they can be very effective when used in extrapolation techniques [2–5]. They can also include superconvergence results which can be very useful in defining effective adaptive strategies. Techniques are being developed [6–8] to combine these ideas with finite difference methods for estimation of higher derivatives in more rigorous error estimation and adaptive methods.

Once a computational result has been obtained, *a posteriori* error estimators and indicators can be utilized to give more specific assessment of errors and to form a basis for many adaptive strategies. *A posteriori* estimators can produce reliable local error assessment judged by effectivity indices. They should be locally computable and can utilize various norms or different error measures, based upon the type of error control desired. There is question as to which *a posteriori* error estimators can be used to produce the most reliable and effective adaptive strategies. Efficiency is the key property of an effective error indicator, especially for large-scale problems.

The efficiencies of a solution method and an associated adaptive process are heavily dependent upon the selection of computer architecture, data structure, and implementation strategy. The domain specification and discretization can be aided via concepts from CAD-CAM. Shephard and colleagues have given important surveys of mesh generation and its associated data structure [9–11]. Concepts of data structure and efficient implementation procedure for time-dependent problems are considered to some extent below.

Although each of the four topics described above is the subject of substantial research in its own right, they should not be considered separately. The strong relationship between each of these aspects of modeling must be investigated. The focus of this paper is on *a* *posteriori* error estimation; however, we show that concepts of implementation and data structure are critical for the effective development and use of error estimators, especially in large-scale dynamic problems.

Given the results of a scientific computation, we want to accomplish two tasks. First, we must be able to assess the accuracy of the computation to see if this level of accuracy meets our design criteria. If it does not, we must define an adaptive strategy to reduce the error in a fixed way. Thus, we state our major goals. Given an initial discrete model with a specified input topology and with boundary and initial conditions that yield a unique solution to our problem, and given an error measure and a specified tolerance, we must choose an adaptive strategy (with error assessment) to ensure that the resulting computational procedure will produce errors satisfying the given tolerance in the given error measure.

In [12], Babuška formulated some of the principle concepts and approaches to adaptivity. He introduced notions of feedback and adaptivity and their relationship to *a posteriori* error estimation and then formulated illustrative theorems in a simple linear, elliptic model setting. He and his colleagues have developed error analyses for one-dimensional problems [13–15]. An important goal is to extend this theory to more complex problems in higher dimensions. These mathematical concepts must form a solid foundation upon which we develop implementation strategies for complex engineering problems.

Given a computational procedure, we define a *feedback mechanism* as a specific set of rules for evaluating the result of the computation, assessing its accuracy in some norm or error measure, and defining a strategy for changing the solution process to improve the results. The feedback mechanism produces a sequence of rules and an associated sequence of approximate solutions $\{\tilde{u}_i\}, i = 1, 2, \cdots$, to the problem. In order to decide when to terminate this process, we need an error estimator $\varepsilon_i(\tilde{u}_i)$; then, the solution $\tilde{u}_i$ is accepted if

$$\varepsilon_i(\tilde{u}_i) \leq \tau \parallel \tilde{u}_i \parallel , \tag{1.1}$$

where τ and $\| \cdot \|$ are *a priori* given tolerances and norms, respectively.

Clearly, there are many possible feedback approaches to any problem. The important concepts are the choice of an error measure, a convergence rate measure, a work measure, and a tolerance. Obviously, these choices are heavily interrelated.

The goal of computation is to obtain the "best" solution possible for a given computational cost. We define an *adaptive process* as a feedback approach which is *optimal* with respect to certain clearly defined objectives. Thus, in order to specify an adaptive process, we must clarify in what sense we mean *optimal* and with respect to which criteria. We could use each of the feedback concepts mentioned above separately as a definition of optimality. Most adaptive processes described in the literature use error and/or convergence rate measures to determine optimality. Although these measures work well for static, elliptic problems, they can produce highly ineffective adaptive strategies for large-scale, time-dependent problems. For these problems, the aspects of work measure such as data structures, available computer architectures, and implementation strategies are critical. We address some of these ideas in Section III.

The use of efficient adaptive strategies for large-scale, time-dependent codes is still in its infancy. Many modern commercial codes incorporate feedback, but very few are truly adaptive with respect to work measures. A detailed, rigorous mathematical analysis of adaptive programs is currently available only for some one-dimensional problems [13–15]. In fact, even the one-dimensional problems are not totally understood. We need to determine mathematical condi‿ions under which the estimators currently being used are valid (e.g. bad mesh spacings, rapidly varying solutions, etc.). We need to determine more robust error estimators for more difficult problems. Also, there are still major differences between the aspects of design criteria and design certification. As will be shown in Section III, there are fundamental differences between adaptive strategies for static and dynamic problems. Finally, information theory currently does not play a significant role in adaptivity, although it has enormous potential.

in Section II, we will present a summary of error estimation and adaptivity techniques and concepts. We also present the major types of refinement or adaptation strategies. In Section III, we discuss the need for efficiency in large-scale, time-dependent problems and discuss recent effective local spatial and temporal adaptive strategies for these problems. We then discuss domain decomposition techniques for de eloping adaptive strategies on large fluid flow problems with various work measure criteria in Section IV.

II. ERROR ESTIMATION AND ADAPTIVITY

Traditional error estimates for finite difference and finite element methods are *a priori* bounds, predicting the asymptotic rate of convergence as the mesh size tends to zero. Unfortunately, this gives us little direct information about the true error for a fixed grid size and approximation space in a difficult problem. However, in many physical applications, there are special features such as wells, nozzles, cracks, corners, obstacles, point loads, etc., which are fixed in location but which greatly affect the solution globally as well as locally. In many of these applications, a special point creates a singularity in the function of interest around the point. *Often*, these singularities are of the form r^{-s} or $log\, r$ where r is the distance to the special point in the domain and $s > 0$ is a real number which gives the strength of the singularity. The *a priori* knowledge of the strength of the singularity can be used very effectively in adaptively grading the grid to take advantage of the asymptotic behavior of the function near the singularity.

We now turn our attention to *a posteriori* error estimates that are based upon information obtained during the solution process. We will distinguish between locally computed *error indicators* and globally valid *error estimators.* This distinction is not uniform in the literature, and many of the *a posteriori* error estimates or estimators are actually local without proof of global validity. Of course, we hope to use indicators to develop estimators due to the computational complexity of global solutions. Full reliability requires global estimation, but important local error assessment can be obtained from local indicators. Adpative improvement algorithms can often be developed using inexpensive error indicators. Essentially all of the work on rigorous *a posteriori* error estimators has been associated with the finite element method. However, various locally computable indicators have been utilized effectively to define adaptive improvement strategies for finite diffference methods. Work is necessary in extending the finite element theories to finite difference methods.

The *a posteriori* error estimators are termed asymptotically correct if the ratio of the estimator to the true error converges to one as the true error tends to zero. Thus, in order to assess the reliability of these estimators, an effectivity index is defined [12,15] as follows:

$$\theta = \frac{\varepsilon}{\| e \|}, \tag{2.1}$$

where ε is the *a posteriori* error estimator and $\| e \|$ is the chosen norm of the error. The estimator is asymptotically correct when $\theta \to 1$ as $\| e \| \to 0$.

For practical application of an erro^r estimator, it is important that $|\theta - 1|$ be small when the error $\| e \|$ is of the order of 10% of the norm of the solution, and should decrease as the error decreases. It is also preferable that we over-estimate the error and that θ be greater than one. The asymptotic correctness of the error estimator is related to superconvergence effects. The mathematical theory of error estimators is discussed in [14–18].

Locally computable *a posteriori* error estimators have been developed primarily by Babuška, Rheinboldt, Bank, Oden, Zienkiewicz, Flaherty, Johnson, and their colleagues [5–8,12,13,15–36]. Excellent surveys of adaptivity and *a posteriori* error estimation have appeared in [37–39]. Under suitable assumptions, their error estimators converge to the norm of the actual error as the mesh size tends to zero. The most recently developed estimators are asymptotically upper bounds for the norm of the true error and can be computed locally, element-by-element (see [40]). These *a posteriori* error estimators are extremely important for problems involving elliptic partial differential equations in determining the reliability of estimates for a fixed grid and a fixed error tolerance in a given norm. The error estimators are used to successively refine locally until the errors in some specific norm are, in some sense, equilibrated. These techniques drive the local refinement at only one or two levels per iteration. Thus, obtaining an "optimal" grid in the sense of equilibrated error in some norm usually takes several iterations. Although the local error estimation is a relatively small part of the solution of an elliptic problem, this is *not* the case for many time-dependent problems with changing local properties. In Section III below, we discuss adaptive techniques which are far more efficient for transient problems.

There are two major types of *a posteriori* error estimators. The first can be described as residual methods. They depend strongly upon the governing operators and thus require care in extensions to more complex nonlinear problems. Although they may require extensive computation to obtain, they can produce effectivity indices that are often quite near to unity. The indicators associated with these methods can be local and fairly inexpensive to compute; the indicators can also help to define effective adaptive improvement algorithms.

The second type of error estimator arises from interpolation techniques. *A priori* estimates and approximation theory play important roles in this class of estimators. Since they are independent of the operator, they can be more general, but often less effective. They generally require *a posteriori* estimation of higher-order derivatives and hence rely heavily upon superconvergence properties. Although these indicators could have poor effectivity indices and thus may be less useful for reliability, the error indicators are cheap and can often be quite useful in developing adaptive improvement techniques.

Depending upon the application, various derivatives of the solution (e.g. buckling loads, stresses, or stress intensity factors) may be as important as or more important than the solution itself in terms of adaptivity and design certification. Thus, different norms or error measures should be used to control different variables. Pointwise estimation of errors is considerably more difficult than the evaluation of local energy norm errors, which is commonly used. Information on higher-order derivatives for higher Sobolev norms can often be obtained via post processing. The analysis and effectiveness of many of these techniques involve concepts of superconvergence and should be combined with *a priori* estimation methods.

Several questions still deserve attention. How should we use error estimators? What can we prove mathematically about the validity of various estimators? Under what conditions are they valid? How can we compare different estimators? Can we use indicators effectively to determine not only where but *how* to improve the process? These are subjects of future research.

When finite element methods are used to discretize model partial differential equations and to approximate the associated functions that have fixed points of singular behavior, there are several ideas for resolving the rapid function growth around the points. Employing these techniques usually amounts to adding more degrees of freedom around the special point to better approximate the rapidly changing function values. One method, commonly called the p-version [14,19,23,28], utilizes higher-order polynomials near the singular points to give better approximation there. Another idea, commonly called the h-version [3,14,19–21], augments the grid around a singular point to obtain better resolution. For certain singularities, the combination, the h-p-version [5,14,32], has been shown to be optimal [12]. Also, nodes can be moved to achieve properties of the h-version with fewer unknowns and hence smaller matrices. The methods [2,26,33,34,41,42] termed the r-version are extremely efficient for one-dimensional problems, but are considerably harder to implement in higher dimensions.

The p-version requires some additional complex code in utilizing the higher-order methods and performing the associated higher-order quadrature techniques. The use of hierarchical basis functions to gradually build up higher-order elements instead of more standard basis functions has been a very efficient way to use the p-version and also somewhat simplifies some of the associated quadratures. The h-version has received more attention for very large problems, partially due to the greater sparsity of the associated matrices, and due to their applicability for finite difference methods.

There are some substantial differences between the theory and application of h-methods and p-methods. In h-methods, $h < 1$ is the parameter and p is the constant. Thus, if θ from (2.1) satisfies

$$\| \theta \| \leq C(p)h^s \tag{2.2}$$

for some $s > 0$, then C will depend strongly upcn p; as $p \to \infty, C \to \infty$. For the p-method, $p > 1$ is the parameter and h is constant. Thus,

$$\| \theta \| \leq C(h)p^{-s} . \tag{2.3}$$

In the h-p-method, both h and p are treated as parameters. In order to obtain a corresponding convergence rate, we let N be the number of degrees of freedom. Then one can show [43] that

$$\| \theta \| \leq C\, exp(-N)^s , \tag{2.4}$$

where $s > 0$ and C is independent of both h and p. In this case, exponential convergence can be obtained, theoretically. These rates have also been achieved in practice for a variety of different applications.

III. STATIC VERSUS DYNAMIC ADAPTIVE IMPROVEMENT

Error estimation and adaptive grid improvement are essential in many different large-scale applications. For specificity, we illustrate the concepts involved via problems related to fluid flow in porous media, although many different applications share the same properties.

Mathematical models of large-scale, time dependent, fluid-flow processes involve large coupled systems of nonlinear evolutionary partial differential equations. In order to compare the results of the models with physical measurements to assess their reliability and to make decisions based on the models, the partial differential equations must be discretized and solved on computers. For example, field-scale simulations of fluid flow in porous media normally involve reservoirs of enormous size although many highly localized phenomena often govern the transport of the fluids. Uniform gridding on the length scale of the local phenomena would involve systems of discrete equations of such great size as to make solution on even the largest computers prohibitive. Therefore, local adaptive grid improvement capabilities and efficient solution processes for the resulting discrete models are becoming more important in reservoir simulation as the fluid-flow processes arising in the applications become more complex and involve more localized phenomena in enormous problems.

In this section, we present examples of localized phenomena, both static and dynamic, that arise in field-scale reservoir simulation. These examples serve to illustrate the difficult problems related to data structures and implementation that arise in many large-scale time-dependent applications. We see the need for efficient computational algorithms designed to take advantage of emerging computer architectures. In the applications described below, we see the dramatic difference between efficient adaptive improvement algorithms for static and dynamic problems. In Section IV, we present domain decomposition techniques which allow adaptive refinement without destroying the efficiency of the underlying large-scale codes and are widely applicable.

For large problems that require the solution of extremely large nonlinear/linear systems of equations, the sparsity structure and banded features of the system matrices can be exploited heavily by special algorithms for present day supercomputers. Many man years have been invested to produce highly vectorized codes that work extremely well for structured (tensor-product) grids for finite difference discretizations. When local improvement strategies were attempted in these codes, the banded structure of the matrices and the corresponding ease of vectorization and efficiency of the codes were destroyed. Although very effective adaptive strategies might be developed with error or convergence rate measures as criteria for optimality, the work measure is greatly increased and the overall efficiency of the codes is reduced if not destroyed. In order to retain the efficiency and reduce the work measure, we must concentrate on data structures and associated solution algorithms that allow for effective use of supercomputer architectures.

One particular data structure, used for local refinement calculations for large-scale reservoir simulation problems, was developed by Uhler, Jones, and the author at Mobil's Field Research Laboratory and is described in [44,45]. It was developed to take advantage of strengths of both the Rheinboldt-Mesztenyi structure [46] and that of Bank and Sherman [25]. This structure, more closely related to that of Rheinboldt and Mesztenyi, has been modified to efficiently support both mesh refinement and the removal of mesh. Neither the Rheinboldt-Mesztenyi nor the Bank-Sherman data structure could allow efficient grid removal; they caused the entire data structure to be generated from scratch for this capability. Thus, although they were extremely effective for elliptic problems, they were inefficient for dynamic applications. The improved data structure supports a grid refinement capability having a set of predetermined macro-cells. Each macro-cell can be locally subdivided after error estimation in an energy norm by a repeated nested refinement into local elements or cells.

The use of refinement at any level of a nested tree structure allows truly local refinement and the equilibration of error in the presence of local phenomena. However, since the length of the branches of the tree can vary greatly, the vectorization of algorithms with this type of data structure is significantly more difficult. There is, however, potential for parallelism in these algorithms [45,47]. Even for static problems where error equilibration has been extremely successful, adaptivity with respect to the work measure requires considerable attention to data structure and efficiency of implementation for large-scale problems.

For truly general local refinement, a complex data structure like that discussed above and the associated complications to the code are necessary. If local refinement is only needed in certain regions or at a very few special points, a technique termed patch refinement may be an attractive alternative. These concepts do not require as complex a data structure but do involve ideas of passing information from one uniform grid to another. Berger and Oliger have been using patch-refinement techniques for hyperbolic problems using finite difference discretizations for some time [48,49]. The local patch-refinement techniques have proven to be very effective in 3-dimensional field-scale petroleum simulations [50,51] for obtaining local resolution around fixed singular points, such as wells, in a reservoir. Although the patch approximation technique is extremely useful in the context of local refinement around a fixed point or region, it can be even more important for dynamic problems.

For time-dependent problems, there is often considerable information which can be used from preceding time steps to help drive the adaptive process. In parabolic problems, where the solution changes smoothly in time, the "optimal" grid used at the previous time step should be a very good approximation to the desired grid at the advanced time step. Thus, beginning with a new coarse grid at each time step and using the elliptic techniques of error estimators to define the local refinement algorithms would be wasteful. For small parabolic problems, when the grid is changing very slowly in time, a much better technique would be to take the grid from the last time step, apply local grid analysis to determine where new grid is needed and where refinement is no longer needed, and then to change only the grid that requires improvement. This requires a data structure that allows efficient removal of the grid as well as grid enhancement. In these techniques, great care must be taken to preserve mass balance when grid is removed and the flow properties must be averaged and described on the new coarser grid.

Although this may be an efficient process for fairly small time-dependent problems, it is generally inefficient to change the grid at each time step for large-scale problems. For large time-dependent problems, iterative processes are generally much more efficient than direct solution techniques. For problems with smoothly changing solutions, the same preconditioner can often be used for several time steps, since the matrices evolve smoothly; this greatly saves in computational effort. If the size of the grid and hence the number of unknowns is constantly changing, clearly the preconditioner must be altered at the same time. Similarly, as mentioned earlier, changing the number of unknowns greatly hinders vectorization techniques. Therefore, a considerably more efficient alternative to constantly adapting the grid is to use a larger refined area within which the action is maintained for several time steps and to move the patch less frequently, after several steps. This idea is similar to the dynamic patch-refinement techniques of Berger and Oliger [48,49].

The adaptivity techniques of Berger and Oliger move the patch approximately every three or four time steps using sophisticated clustering techniques. Within the patches, Richardson-type methods are used to estimate the local truncation error to decide where grid refinement is needed. The error estimators are independent of both the partial differential equations and the difference methods used to discretize them. These methods have considerable potential for use in large-scale fluid-flow problems in addition to the hyperbolic problems to which they have been applied.

For hyperbolic or advection-dominated parabolic partial differential equations arising in fluid-flow problems, sharp fluid interfaces move along characteristic or near-characteristic directions. The computed fluid velocities determine both the local speed and direction of the fronts, identifying the regions where local refinement will be needed at the upcoming time steps. This information should be utilized in the adaptive method to move the local refinement with the front. We are currently experimenting with using the computed fluid velocities to move the patch grids in characteristic directions in quantum jumps. Thus, *a posteriori* information is being used to determine *how* to adapt the grid in an effective manner.

The patch refinement techniques described above have been used to follow fluid interfaces in multiphase problems in [52–55]. These methods utilize a coarse grid to define fluid velocities which are then combined with a modified method of characteristics and local, patch refinement around the moving fronts.

Another problem that has plagued large-scale reservoir simulation is the difficulty in treating local transients around wells in fully implicit codes. When one well is rapidly opened or shut to flow, the local fluid properties around the well change sufficiently quickly that the global Newton-Raphson method used to linearize the flow equations does not converge. The present industrial solution is to cut the time step over the whole reservoir to obtain convergence, even though the difficulty is highly localized. This is extremely wasteful, and computationally intensive. The domain decomposition techniques discussed in the next section have been used to develop accurate and efficient local time-stepping algorithms [56,57] and to obtain local initial guesses for the Newton-Raphson iteration.

The applications described above indicate that for large-scale dynamic problems, efficiency of implementation is the key to adaptivity, and major advances have been made in this area. We cannot rely only upon error or rate convergence measures but must consider computational costs. Techniques based on domain decomposition ideas give us efficient computational procedures. Therefore, the work measure can be used more effectively with these algorithms due to their superior computational efficiency.

IV. ADAPTIVE IMPROVEMENT VIA DOMAIN DECOMPOSITION TECHNIQUES

In large-scale simulation problems, attempts to implement local grid refinement can often destroy the efficiency of existing codes. In this section, we describe patch refinement methods that can easily be incorporated into large existing codes without seriously affecting their efficiency. The methods discussed maintain accuracy of the discretization across patch interfaces and can also take full advantage of the parallel and vector capabilities of the emerging supercomputers. These techniques are related to various domain decomposition methods [52–63]. High accuracy is obtained throughout the computational region by incorporating local refinements in patches wherever they are needed. A composite grid is obtained by superimposing these refinements on a quasi-uniform grid on the original domain. Previous techniques usually have no systematic way of dealing with such questions as interface interpolation, mass conservation, and degree of grid overlap, and also usually involve the solution of the coarse-grid problems with the regions corresponding to the refinement removed. This destroys the banded structure and ease of vectorization of the matrices for coarse-grid regions.

In the methods discussed below, the problem is formulated with a composite operator on the composite grid. The techniques are iterative procedures which drive the residual of this composite-grid operator toward zero. Composite-grid operators for the finite element discretization were derived in [58,62,63]. Two similar composite-grid methods are the FAC method [59,60] and the BEPS method described in [52,58,61,62]. Both methods can be described as preconditioners for other iterative methods [58,60] or as full iterative methods [59,61]. When the methods are used for point-centered finite differences or finite element methods which satisfy a variational principle, they do not require a scaling as an iterative method; if cell-centered finite difference methods are used, the analysis is considerably more complex and a scaling may be necessary for rapid convergence [61,62]. Complete error analyses for finite-difference-based composite-grid operators for variable coefficients appear in [62]. Local time-stepping methods based upon these same techniques will appear soon [56,57]. Use of local refinement in conjunction with mixed finite element methods has been described in [57,63].

Essentially all of the concepts for the applications described above can be related to a general algorithm and described in simple algebraic terms. See [64] for a more detailed description. We outline the procedure by considering the following discrete problem on a composite grid ω:

$$\mathbf{A}\mathbf{y} = \mathbf{b} .$$

We solve this problem using a preconditioned conjugate gradient method. Here, we present two methods for constructing a preconditioning matrix $\mathbf{B}$.

We first introduce some notations. We let Ω be our computational domain and consider two grids on Ω. We introduce the matrix $\tilde{\mathbf{A}}$ corresponding to an approximation of our problem on a regular, quasi-uniform grid $\tilde{\omega}$ on Ω. Let $\Omega_2 \subset \Omega$ be a region that contains some local phenomenon that may require better approximation. Let $\Omega_1 = \Omega \setminus \Omega_2$. Then Ω_1 and Ω_2 produce a natural decomposition of Ω. The composite grid ω has the same grid as $\tilde{\omega}$ in Ω_1, denoted by $\tilde{\omega}_1$, but has refined grid in Ω_2, denoted by ω_2; $\tilde{\omega}_2$ denotes the coarse-grid points in Ω_2. If $\mathbf{y}$ and $\tilde{\mathbf{y}}$ are partitioned into $\mathbf{y} = \binom{\mathbf{y}_1}{\mathbf{y}_2}$ and $\tilde{\mathbf{y}} = \binom{\tilde{\mathbf{y}}_1}{\tilde{\mathbf{y}}_2}$, then this induces corresponding partitions for $\mathbf{A}$ and $\tilde{\mathbf{A}}$:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad \tilde{\mathbf{A}} = \begin{pmatrix} \tilde{\mathbf{A}}_{11} & \tilde{\mathbf{A}}_{12} \\ \tilde{\mathbf{A}}_{21} & \tilde{\mathbf{A}}_{22} \end{pmatrix}.$$

Both the FAC-preconditioner [59,60] and the BEPS-preconditioner (Bramble, Ewing, Pasciak, and Schatz [58]) follow three common steps. The common steps in the algorithm for solving $\mathbf{B}\mathbf{y} = \mathbf{b}$ are:

(i) solve the fine-grid problem in Ω_2

$$\mathbf{A}_{11}\mathbf{y}_1^F = \mathbf{b}_1 \;;$$

(ii) restrict the defect on the coarse grid

$$\tilde{\mathbf{d}} = \mathbf{P}^T\left(\mathbf{b} - \mathbf{A}\begin{pmatrix}\mathbf{y}_1^F \\ 0\end{pmatrix}\right) = \mathbf{P}^T\begin{pmatrix} 0 \\ \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1\end{pmatrix}$$

($\mathbf{P}^T$ is the restriction operator);

(iii) solve for the coarse-grid correction

$$\tilde{\mathbf{A}}\tilde{\mathbf{c}} = \tilde{\mathbf{d}} \,.$$

The FAC and BEPS methods differ in the last step:

(iv) (FAC) interpolate this correction over the fine grid in Ω_2

$$\mathbf{c} = \mathbf{P}\tilde{\mathbf{c}} \;;$$

then,

$$\mathbf{y} = \mathbf{B}^{-1}\mathbf{b} = \begin{pmatrix}\mathbf{y}_1^F \\ 0\end{pmatrix} + \mathbf{c} \,.$$

(iv) (BEPS) find the harmonic component $\mathbf{y}_1^H$ on the fine grid

$$\mathbf{A}_{11}\mathbf{y}_1^H = \mathbf{A}_{12}\tilde{\mathbf{c}}_2, \qquad \text{where} \qquad \tilde{\mathbf{c}} = \begin{pmatrix} \tilde{\mathbf{c}}_1 \\ \tilde{\mathbf{c}}_2 \end{pmatrix};$$

then,

$$\mathbf{y} = \mathbf{B}^{-1}\mathbf{b} = \begin{pmatrix} \mathbf{y}_1^F + \mathbf{y}_1^H \\ \tilde{\mathbf{c}}_2 \end{pmatrix}.$$

An important feature of these methods is that both preconditioning matrices **B** are spectrally equivalent to the composite-grid matrix **A** with constants that do not depend on the mesh size [62]. Therefore, the preconditioned iterative procedures based upon gradient-type methods are optimal.

These two methods have one common step—solving the coarse-grid problem, which (we assume) can be done very efficiently by some fast solver or within the technology of the existing code maintaining a high level of vectorization and/or parallelization. FAC and BEPS differ on one important issue: on the last step, BEPS solves one more problem on the fine grid (harmonic component), securing in this way the symmetry of the preconditioning matrix **B**. Instead of that, FAC interpolates the coarse-grid correction and adds it to the fine-grid component. Thus, in general, in FAC we solve one fine-grid problem less, but we use interpolation. A restriction is that the interpolation operator **P** should be equal to the transposed restriction operator (from fine to coarse grid); moreover, the composite-grid matrix **A** and coarse-grid matrix $\tilde{\mathbf{A}}$ should satisfy the relation $\mathbf{A} = \mathbf{P}\tilde{\mathbf{A}}\mathbf{P}^T$ (so-called variational condition) [59]. This is automatically satisfied only for finite element problems. The FAC also produces a nonsymmetric preconditioning matrix **B**, and generalizations of the standard CG-methods that do not take advantage of conjugacy should be used. However, as has been shown in [62], this matrix is symmetric in a certain subspace. Performing the iteration within this subspace will lead exactly to the last local problem of the BEPS preconditioner. See [64] for more details.

The domain decomposition concept mentioned above involves the development of a preconditioner. This preconditioner is novel in that the task of computing its inverse applied to a vector reduces to the solution of separate matrix systems for the local refinements and the matrix system for the quasi-uniform grid on the original domain. Note that this quasi-uniform grid overlaps the regions of local refinement, and that its corresponding matrix problem remains invariant when local refinements are dynamically added or removed. This local refinement technique can be incorporated in existing reservoir codes without extensive modification. See [50,51] for examples of implementation of these methods in an industrial code. Furthermore, if the nodes on the quasi-uniform grid are chosen in a regular pattern, highly vectorizable algorithms for the solution of the corresponding matrix system can be developed.

In conclusion, we see that the domain decomposition techniques have enormous potential for greatly reducing computational complexity of codes in connection with adaptive improvement strategies. They allow realistic adaptivity with respect to work measure by addressing efficient implementation Then the domain decomposition methods can be combined with characteristic flow indicators as in [53,54] to determine how to adapt the grid for dynamic moving front problems—a severe challenge to any adaptive method.

ACKNOWLEDGEMENTS

The research reported here was partly supported by Office of Naval Research Grant No. 0014-88-K-0370, and by the Institute for Scientific Computation at the University of

Wyoming through National Science Foundation Grant No. RII-8610680.

REFERENCES

[1] I. Babuška, Adaptive mathematical modeling, in: J. E. Flaherty, P. J. Paslow, M. S. Shephard and J. D. Vasilakis, eds., Adaptive Methods for Partial Diferential Equations, (SIAM, Philadelphia, Pennsylvania 1989) 1–14.

[2] A. R. Diaz, N. Kikuchi and J. E. Taylor, A method of grid optimization for finite element methods, Comp. Meth. Appl. Mech. and Engrg. 41 (1983) 29–45.

[3] L. Demkowicz, J. T. Oden and P. Devloo, An h-type mesh refinement strategy based on a minimization of interpolation error, Comp. Meth. Appl. Mech. and Engrg. 53 (1985) 67–89.

[4] L. Demkowicz, and J. T. Oden, On a mesh optimization method based on a minimization of interpolation error, Int. J. Eng. Sci. 24 (1986) 55–68

[5] J. T. Oden, L. Demkowicz, L. Westermann and W. Rachowicz, Toward a universal $h-p$ adaptive finite element strategy, 2: A posteriori error estimates, Comp. Meth. Appl. Mech. and Engrg., to ›pear.

[6] K. Eriksson and C. Johnson, An adaptive method for linear elliptic problems, Technical Report 1985-13, Math. Dept., Chalmers University of Technology; and Math. Comp., to appear.

[7] K. Eriksson, C. Johnson and J. Lennblad, Error estimates and automatic time and space step control for linear parabolic problems, Technical Report, 1986, Math. Dept., Chalmers University of Technology; and SIAM J. Numer. Anal., to appear.

[8] C. Johnson, Y. Y. Nie and V. Thomée, An a posteriori estimate and automatic time step control for a backward Euler discretization of a parabolic problem, Technical Report 1985-23, Math. Dept., Chalmers University of Technology; and SIAM J. Numer. Anal., to appear.

[9] M. S. Shephard, Approaches to the automatic generation and control of finite element meshes, Applied Mechanics Review 41 (1988) 169–185.

[10] M. S. Shephard, P. L. Baehmann and K. R.Grice, The versatility of automatic mesh generators based on tree structures and advanced geometric constructs, Communications in Applied Numerical Methods 4 (1988) 379–392.

[11] P. L. Bachmann, S. L. Wittchen, M. S. Shephard, K. R. Grice and M. A. Yerry, Robust geometrically based automatic two-dimensional mesh generation, Int. J. Num. Math. Energy 24 (1987) 1043-1078.

[12] I. Babuška, Feedback, adaptivity and a posteriori estimates in finite elements: aims, theory, and experience, in: I. Babuška, O. C. Zienkiewicz, J. Gago and E. R. de A. Oliveria, eds., Accuracy Estimates and Adaptive Refinements in Finite Element Computations (Wiley, New York, 1986).

[13] I. Babuška and M. Vogelius, Feedback and adaptive finite element solution of one-dimensional boundary value problems, Numerische Mathematik 44 (1984) 75–102.

[14] W. Gui and I. Babuška, The h, p, and $h-p$ versions of the finite element method in one dimension, 1: The error analysis of the p-version, 2: The error analysis of the h and

h–p versions, 3: The adaptive *h–p* version, Numerische Mathematik 48 (1986) 557–612, 613-657, 658-683.

[15] I. Babuška and W. C. Rheinboldt, A posteriori error analysis of finite element solutions for one-dimensional problems, SIAM J. Num. Anal. 18 (1981), 565–589.

[16] I. Babuška and A. Miller, A posteriori error estimate and adaptive techniques for the finite element method, Technical Note BN-968, Institute for Physical Science and Technology, University of Maryland (1981).

[17] I. Babuška and D. Yu, Asymptotically exact a posteriori error estimator for biquadratic elements, Technical Note BN-1050, Institute for Physical Science and Technology, University of Maryland (1986).

[18] R. E. Bank and A. Weiser, Some a posteriori error estimates for elliptic partial differential equations, Math. Comp. 44 (1985) 283–301.

[19] I. Babuška and M. R. Dorr, Error estimates for the combined *h* and *p* versions of the finite element method, Technical Report BN-951, Institute for Physical Sciences and Technology, University of Maryland, College Park, Maryland (1980).

[20] I. Babuška and W. C. Rheinboldt, A-posteriori error estimates for the finite element method, Internat. J. Num. Methods Engrg. 12 (1978) 1597–1615.

[21] I. Babuška and W. C. Rheinboldt, Reliable error estimation and mesh adaptation for the finite element method, in: J. T. Oden, ed., Computational Methods in Nonlinear Mechanics (North-Holland, New York, 1984).

[22] I. Babuška and W. C. Rheinboldt, A survey of a-posteriori error estimators and adaptive approach in the finite element method, Technical Report BN-1981, Laboratory for

Numerical Analysis, University of Maryland, College Park, Maryland (1982).

[23] I. Babuška, B. A. Szabó and I. N. Katz, The p-version of the finite element method, SIAM J. Numer. Anal. 18 (1981) 515–545.

[24] R. E. Bank, Locally computed error estimates for elliptic equations, in: E. R. Arantes e Oliveira, I. Babuška, O. C. Zienkiewicz and J. P. Gago, eds., Proceedings International Conference on Accuracy Estimates and Adaptive Refinements in Finite Element Computations, Vol. 1, Libson, Portugal (1981) 21–30.

[25] R. E. Bank and A. H. Sherman, PLTMG users' guide, Technical Report No. 152, Center for Numerical Analysis, University of Texas at Austin (1979).

[26] J. T. Oden, L. Demkowicz and T. Strouboulis, Adaptive finite element methods for flow problems with moving boundaries. I: Variational principles and a posteriori estimates, Comp. Meth. Appl. Mech. and Engrg. 46 (1984) 217–251.

[27] J. T. Oden, T. Strouboules, P. Devloo and M. Howe, Recent advances in error estimation and adaptive improvement of finite element calculations, in: A. K. Noor, ed., Computational Mechanics Advances and Trends, Vol. 75 (ASME 1986) 369–400.

[28] O. C. Zienkiewicz and A. W. Craig, Adaptive mesh refinement and a posteriori error estimation for the *p*-version of the finite element method, in: I. Babuška, J. Chandra and J. E. Flaherty, eds., Adaptive Computational Methods for Partial Differential Equations (SIAM, Philadelphia, Pennsylvania, 1983) 33–57.

[29] O. C. Zienkiewiez and J. Z. Zhu, A simple error estimator and adaptive procedure for practical engineering analysis, Int. J. Num. Meth. Eng. 24 (1987) 337–357.

[30] O. C. Zienkiewicz, J. Z. Zhu, Y. C. Liu, K. Morgan and J. Peraire, Error estimates

and adaptivity from elasticity to high speed compressible flow, in: J. R. Whiteman. ed., MAFELAP VI (Academic Press, London, 1988) 483–512.

[31] M. Ainsworth, J. Z. Zhu, A. W. Craig and O. C. Zienkiewicz, Analysis of the Zienkiewicz-Zhu a posteriori error estimator in the finite element method, Int. J. Num. Math. Eng., to appear.

[32] O. C. Zienkiewicz, J. Z. Zhu, A. W. Craig and M. Ainsworth, Simple and practical error estimation and adaptivity: h and $h-p$ version procedures, in: J. E. Flaherty, P. J. Paslow, M. S. Shephard and J. D. Vasilakis, eds., Adaptive Methods for Partial Differential Equations (SIAM, Philadelphia, Pennsylvania, 1989) 100–114.

[33] S. Adjerid and J. E. Flaherty, A moving finite element method with error estimation and refinement for one-dimensional time dependent partial differential equations, SIAM J. Numer. Anal. 23 (1986) 779–796.

[34] D. C. Arney and J. E. Flaherty, A two-dimensional mesh moving technique for time-dependent partial differential equations, J. Comp. Phys. 67 (1986) 124–144.

[35] J. M. Coyle, J. E. Flaherty and R. Ludwig, On the stability of mesh equidistribution strategies for time-dependent partial differential equations, J. Comp. Phys. 62 (1986) 26–39.

[36] S. Adjerid and J. E. Flaherty, Second-order finite element approximations and a posteriori error estimation for two-dimensional parabolic systems, Numer. Math 53 (1988) 183–198.

[37] A. K. Noor and I. Babuška, Quality assessment and control of finite element solutions, Finite Elements in Analysis and Design 3 (1987) 1–26.

[38] J. T. Oden and L. Demkowicz, Advances in adaptive improvements: a survey of adaptive methods in computational fluid mechanics, in: A. K. Noor and J. T. Oden, eds., State of the Art Surveys in Computational Mechanics (American Society of Mechanical Engineers, New York, 1988).

[39] J. T. Oden, Notes on a posteriori error estimates for finite element approximations of boundary and initial-value problems, in preparation.

[40] A. Weiser, Local-mesh, local-order, adaptive finite element methods with a posteriori error estimates for elliptic partial differential equations, Technical Report No. 213, Department of Computer Science, Yale University, New Haven, Connecticut (1981).

[41] K. Miller and R. N. Miller, Moving finite elements, I, II, SIAM J. on Numerical Analysis 18 (1981) 1019–1032, 1033–1057.

[42] L. Demkowicz and J. T. Oden, An adaptive characterisitic Petrov-Galerkin finite element method for convection-dominated linear and nonlinear problems in two space variables, Comp. Meth. in Appl. Mechanics and Engng. 55 (1986) 63–87.

[43] B. Guo and I. Babuška, The $h-p$ version of the finite element method: I, J. Comp. Mechanics 1 (1986) 21–42.

[44] J. C. Diaz, R. E. Ewing, R. W. Jones, A. E. McDonald, L. M. Uhler and D. U. von Rosenberg, Self-adaptive local grid refinement for time-dependent, two-dimensional simulation, in: G. F. Carey, R. H. Gallagher, J. T. Oden and O. C. Zienkiewicz, eds., Finite Elements in Fluids, Vol. VI (Wiley, New York, 1985) 279–290.

[45] J. C. Diaz and R. E. Ewing, Potential of HEP-like MIMD architectures in self adaptive local grid refinement for accurate simulation of physical processes, Proceedings Workshop on Parallel Processing Using the HEP, Norman, Oklahoma (1985).

[46] W. C. Rheinboldt and C. K. Mesztenyi, On a data structure for adaptive finite element mesh refinement, Trans. Math. Software 6 (1980) 166–187.

[47] C. G. Macedo, Jr., J. C. Diaz and R. E. Ewing, A knowledge-based system for the determination of activity indicators for self-adpative grid methods, Mathematics and Computers in Simulation 32 (1989) 431–439 and in: E.Houstis, J. Rice, R. Vichnevetsky, eds., Fourth Generation Mathematical Software Systems, to appear.

[48] M. J. Berger, Data structures for adaptive mesh refinement, in: I. Babuška, J. Chandra and J. E. Flaherty, eds., Adaptive Computational Methods for Partial Differential Equations (SIAM, Philadelphia, Pennsylvania, 1983) 237–251.

[49] M. J. Berger and J. Oliger, Adaptive mesh refinement for hyperbolic partial differential equations, Man. NA-83-02, Computer Science Department, Stanford University, California (1983).

[50] R. E. Ewing, B. A. Boyett, D. K Babu and R. F. Heinemann, Efficient use of locally refined grids for multiphase reservoir simulation, SPE 18413, Proceedings Tenth SPE Symposium on Reservoir Simulation, Houston, Texas (1989) 55–70.

[51] R. E. Ewing, B. A. Boyett and M. S. El-Mandouh, Local grid refinement for reservoir simulation, Proceedings of the SIAM Conference on Mathematical and Computational Issues in Geophysical Fluid and Solid Mechanics (SIAM, Philadelphia, Pennsylvania), submitted.

[52] M. Espedal and R. E. Ewing, Characteristic Petrov-Galerkin subdomain methods for two-phase immiscible flow, Comp. Meth. in Appl. Mechanics and Engrg. 64 (1987) 113–135.

[53] H. K. Dahle, M. S. Espedal and R. E. Ewing, Characteristic Petrov-Galerkin subdomain methods for convection diffusion problems, in: M. F. Wheeler, ed., IMA Volume 11, Numerical Simulation in Oil Recovery (Springer Verlag, Berlin, 1988) 77–88.

[54] H. K. Dahle, M. S. Espedal, R. E. Ewing and O. Saevareid, Characteristic adaptive subdomain methods for reservoir flow problems, Numerical Solutions of Partial Differential Equations, to appear.

[55] R. E. Ewing, M. S. Espedal, T. F. Russell and O. Saevareid, Reservoir simulation using mixed methods, a modified method of characteristics, and local grid refinement, Proceedings of Joint IMA/SPE European Conference on The Mathematics of Oil Recovery, Robinson College, Cambridge University, 1989, to appear.

[56] R. E. Ewing, R. D. Lazarov and P. S. Vassilevski, Finite difference schemes for parabolic problems on grids with local refinement in time and in space, Computing, submitted.

[57] R. E. Ewing, P. Jacobs, R. Parashkevov and J. Shen, Applications of adaptive grid refinement methods, Proceedings of Fifth IIMAS Workshop on Numerical Analysis, (SIAM, Philadelphia, Pennsylvania), to appear.

[58] J. Bramble, R. E. Ewing, J. Pasciak and A. Schatz, A preconditioning technique for the efficient solution of problems with local grid refinement, Comp. Meth. in Appl. Mechanics and Engrg. 67 (1988) 149–159.

[59] S. McCormick and J. Thomas, The fast adaptive composite grid method for elliptic boundary value problems, Math. Comp. 46 (1986) 439–456.

[60] J. Mandel and S. McCormick, Iterative solution of elliptic equations with refinements: the two-level case, Proceedings Second International Symposium on Domain Decomposition Methods, Los Angeles, California, 1988, to appear.

[61] R. E. Ewing, Domain decomposition techniques for efficient adaptive local grid refinement, ISC Report #1988-04 and in: T. F. Chan, R. Glowinski, J. Periaux, O. B. Widlund, eds., Domain Decomposition Methods (SIAM, Philadelphia, Pennsylvania, 1989) 192–206.

[62] R. E. Ewing, R. D. Lazarov and P. S. Vassilevski, Local refinement techniques for elliptic problems on cell-centered grids, I: Error Analysis, Math Comp., submitted.

[63] R. E. Ewing, R. D. Lazarov, T. F. Russell and P. S. Vassilevski, Local refinement via domain decomposition techniques for mixed finite element methods with rectangular Raviart-Thomas elements, Proceedings of 1989 Conference on Domain Decomposition Methods (SIAM Publications), to appear.

[64] R. E. Ewing, Application of domain decomposition techniques in large-scale fluid flow problems, Applied Numerical Mathematics, to appear.

PARALLEL COMPUTATION WITH ADAPTIVE METHODS

FOR ELLIPTIC AND HYPERBOLIC SYSTEMS*

Messaoud BENANTAR, Rupak BISWAS, Joseph E. FLAHERTY

Department of Computer Science, Rensselaer Polytechnic Institute,

Troy, NY 12180, USA

Mark S. SHEPHARD

Rensselaer Design Research Center, Rensselaer Polytechnic Institute,

Troy, NY 12180, USA

ABSTRACT

We consider the solution of two-dimensional vector systems of elliptic and hyperbolic partial differential equations on a shared memory parallel computer. For elliptic problems, the spatial domain is discretized using a finite quadtree mesh generation procedure and the differential system is discretized by a finite element-Galerkin technique with a piecewise linear polynomial basis. Resulting linear algebraic systems are solved using the conjugate gradient technique with element-by-element and symmetric successive over-relaxation preconditioners. Stiffness matrix assembly and linear system solutions are processed in parallel with computations scheduled on noncontiguous quadrants of the tree in order to minimize process synchronization. Determining noncontiguous regions by coloring the regular finite quadtree structure is far simpler than coloring elements of the unstructured mesh that the finite quadtree procedure generates. We describe linear-time complexity coloring procedures that use six and eight colors.

For hyperbolic problems, the rectangular spatial domain is discretized into a grid of rectangular cells, and the differential system is discretized by

* This research was partially supported by the U. S. Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant Number AFOSR 85-0156; by the SDIO/IST under management of the U. S. Army Research Office under Contract Number DAAL 03-86-K-0112; and by the National Science Foundation under Institutional Infrastructure Grant Number CDA-8805910.

an explicit finite difference technique. Recursive local refinement of the time steps and spatial cells of a coarse base mesh is performed in regions where a refinement indicator exceeds a prescribed tolerance. Data management involves the use of a tree of grids with finer grids regarded as offspring of coarser ones. Computational procedures that sequentially traverse the tree structure while processing solutions on each grid in parallel and that process solutions at the same tree level in parallel have been developed. Computational results using the sequential tree traversal scheme are presented and compared with results using a non-adaptive strategy. Heuristic processor load balancing techniques are suggested for the parallel tree traversal procedure.

1. Introduction

Computational demands of engineers and scientists have been one of the principal driving forces in the development of increasingly powerful digital computers. High processing speed is essential to applications such as real-time simulation, signal processing, and systems involving partial differential equations. Computational solutions of existing and envisioned problems can easily require several hours to days on the most powerful uniprocessor computers.

Conventional fixed-step and fixed-order finite difference and finite element techniques for solving partial differential equations are therefore giving way to adaptive solution techniques, which generally provide much greater efficiency on complex multi-dimensional problems. Current adaptive strategies are classified as h-, p-, or r-refinement when, respectively, computational meshes are refined or coarsened in regions where more or less resolution is needed [1, 2], the order of accuracy is varied in different regions [3], or when a mesh of fixed topology is moved to follow dynamic phenomena [4]. Combinations of hr- or hp-refinement straiegies have also been applied to parabolic [5-8], hyperbolic [1, 9], and elliptic probiems [10]. The particular combination of h- and p-refinement has been shown to yield exponential convergence rates in certain situations [10]. Adaptive software typically includes techniques for constructing a posteriori estimates of local or global discretization errors [6, 11] and such reliability measures must no longer De ignored.

Advances in adaptive software and methodology notwithstanding, parallel computational strategies will be an essential ingredient in the solution of larger and more complex problems. Models of parallel computation are based on shared memory and distributed memory architectures. Distributed memory systems tend to have large numbers of relatively simple processing elements connected in a network. Available memory on these fine-grained systems is distributed with the processing elements at the nodes of the network, so data access is by message passing. Systems built using hypercube-, butterfly-, and grid-connected networks are all commercially available. Balancing communication and synchronizing processing is extremely important because processing elements are typically operating in lock-step fashion in order to improve throughput and processor utilization. Mapping specific applications to a particular network can be difficult, particularly with adaptive methods that use hierarchical data structures. A poor mapping will require extensive global communication which is expensive relative to network routing.

Shared memory systems involve a more coarse-grained level of parallelism with relatively few processors operating asynchronously and communicating with a global memory, although variations are common. For example, processing elements may have a local cache memory in order to reduce, for instance, bus contention and may have vector capabilities; thus, providing a hierarchy of coarse- and fine-grained parallelism.

Our goal is to develop parallel adaptive methods for partial differential equations. At this juncture, we have been experimenting with algorithms for finite element solutions of elliptic problems and explicit finite difference solutions of hyperbolic problems on shared-memory computers. As described in Section 2, linear self-adjoint elliptic problems are discretized using a piecewise linear polynomial basis on a grid of triangular elements that are obtained by finite quadtree mesh generation (cf. Section 2.1). Resulting linear algebraic systems are solved by preconditioned conjugate gradient iteration using either an element-by-element or symmetric successive over-relaxation preconditioning (cf. Section 2.2). Parallel processing of the stiffness matrix assembly and linear system solutions on noncontiguous regions reduces process synchronization. Coloring the more regular finite quadtree structure is a much simpler means of determining noncontiguous regions than coloring the unstructured mesh that the finite quadtree procedure generates. In Section 2.3, we describe linear-time complexity quadtree coloring procedures that use six and eight colors. Two simple examples, presented in Section 2.4, demonstrate the high degree of parallelism that is possible using this approach.

Our research on a parallel adaptive technique for hyperbolic systems, based on a serial algorithm of Arney and Flaherty [1], is described in Section 3. Recursive local refinement of the cells of a coarse space-time mesh is performed in regions of high discretization error. Data management involves the use of a tree of grids with finer grids regarded as offspring of coarser ones. Computational procedures may either sequentially traverse the tree structure while processing solutions on each grid in parallel (cf. Section 3.1) or process solutions at the same tree level in parallel (cf. Section 3.3). Computational results using the sequential tree traversal scheme, presented in Section 3.2, are compared with results using a non-adaptive strategy. Conclusions and future considerations are discussed in Section 4.

2. Elliptic Problems

Consider a two-dimensional linear elliptic problem in m variables having the form

$$-[\mathbf{D}_1(x,y)\mathbf{u}_x]_x - [\mathbf{D}_2(x,y)\mathbf{u}_y]_y + \mathbf{Q}(x,y)\mathbf{u} = \mathbf{f}(x,y), \quad (x,y)\in\Omega, \tag{2.1a}$$

$$u_i(x,y) = g_i(x,y), \quad (x,y)\in\partial\Omega_i^D, \quad i = 1, 2, \cdots, m, \tag{2.1b}$$

$$[\mathbf{D}_1(x,y)\mathbf{u}_x n_1 + \mathbf{D}_2(x,y)\mathbf{u}_y n_2]_i = g_i(x,y), \quad (x,y)\in\partial\Omega_i^N, \quad i = 1, 2, ..., m, \tag{2.1c}$$

where $\mathbf{D}_1(x,y)$, and $\mathbf{D}_2(x,y)$ are symmetric positive definite $m \times m$ matrices; $\mathbf{Q}(x,y)$ is a symmetric positive semi-definite $m \times m$ matrix; $\mathbf{u}(x,y)$ and $\mathbf{f}(x,y)$ are m-vectors; $\partial\Omega = \partial\Omega_i^D \bigcup \partial\Omega_i^N$, $i = 1, 2, \cdots, m$, is the boundary of the bounded domain Ω; and $\mathbf{n} = [n_1, n_2]^T$ is a unit outer normal to $\partial\Omega$.

The Galerkin form of (2.1) consists of determining $\mathbf{u} \in H_E^1$ satisfying

$$A(\mathbf{v},\mathbf{u}) + (\mathbf{v},\mathbf{f}) = \sum_{i=1}^{m} \int_{\partial\Omega_i^N} v_i g_i \, ds, \quad \text{for } all \ \mathbf{v} \in H_0^1, \tag{2.2a}$$

where

$$A(\mathbf{v},\mathbf{u}) = \int_\Omega [\mathbf{v}_x^T \mathbf{D}_1 \mathbf{u}_x + \mathbf{v}_y^T \mathbf{D}_2 \mathbf{u}_y + \mathbf{v}^T \mathbf{Q} \mathbf{u}] \, dxdy, \quad (\mathbf{v},\mathbf{u}) = \int_\Omega \mathbf{v}^T \mathbf{u} \, dxdy. \tag{2.2b,c}$$

As usual, the Sobolev space H^1 consists of functions having first partial derivatives in L^2. The subscripts E and 0 further restrict functions to satisfy the essential boundary conditions (2.1b) and trivial versions of (2.1c), respectively. Finite element solutions of (2.2) are constructed by approximating H^1 by a finite dimensional subspace S^N and determining $\mathbf{U} \in S_E^N$ such that

$$A(\mathbf{V},\mathbf{U}) + (\mathbf{V},\mathbf{f}) = \sum_{i=1}^{m} \int_{\partial\Omega_i^N} V_i g_i \, ds, \quad \text{for } all \ \mathbf{V} \in S_0^N. \tag{2.3}$$

Selecting S^N as a space of continuous piecewise linear polynomials with respect to a partition of Ω into triangular finite elements (cf. Section 2.1), substituting these approximations into (2.3), and evaluating the integrals by quadrature rules yields a sparse, symmetric, positive definite, N-dimensional linear system of the form

$$\mathbf{KX} = \mathbf{b}, \tag{2.4}$$

where $\mathbf{X}$ is an N-vector of Galerkin coordinates.

2.1. Finite Quadtree Mesh Structure

Meshes of triangular or quadrilateral elements are created automatically on Ω by using the *finite quadtree* procedure [12]. With this technique, Ω is embedded in a square "universe" that may be recursively quartered to create a set of disjoint squares called *quadrants*. Data associated with quadrants is managed using a hierarchical tree structure with the original square universe regarded as the root and with smaller quadrants created by subdivision regarded as offspring of larger ones. Quadrants intersecting $\partial\Omega$ are recursively quartered until a prescribed spatial resolution of Ω has been obtained. At this stage, quadrants that are leaf nodes of the tree and intersect $\Omega \cup \partial\Omega$ are further divided into small sets of triangular or quadrilateral elements. Severe mesh gradation is avoided by imposing a maximal one-level difference between quadrants sharing a common edge. This implies a maximal two-level difference between quadrants sharing a common vertex. A final "smoothing" of the triangular or quadrilateral mesh improves element shapes and further reduces mesh gradation near $\partial\Omega$.

A simple example involving a domain consisting of a rectangle and a quarter circle, as shown in Figure 1, will illustrate the finite quadtree process. In the upper left portion of the figure, the square universe containing the problem domain is quartered creating the one-level tree structure shown at the upper right. Were this deemed to be satisfactory geometrical resolution, a mesh of five triangles could be created. As shown, the triangular elements are associated with quadrants of the tree structure. In the example shown in the lower portion of Figure 1, the quadrant containing the circular arc is quartered and the resulting quadrant that intersects the circular arc is quartered again to create the three-level tree shown in the lower right portion of the figure. A triangular mesh generated on this tree structure is also shown. Mesh smoothing, that normally follows element creation, is not shown.

Arbitrarily complex two-dimensional problem domains may be discretized in this manner and generally produce unstructured grids; however, the underlying tree of quadrants remains regular. Further solution-based mesh refinement is easily accomplished by subdividing appropriate leaf-node quadrants and generating a new mesh of triangular or quadrilateral elements locally; thus, unifying the mesh generation and adaptive solution

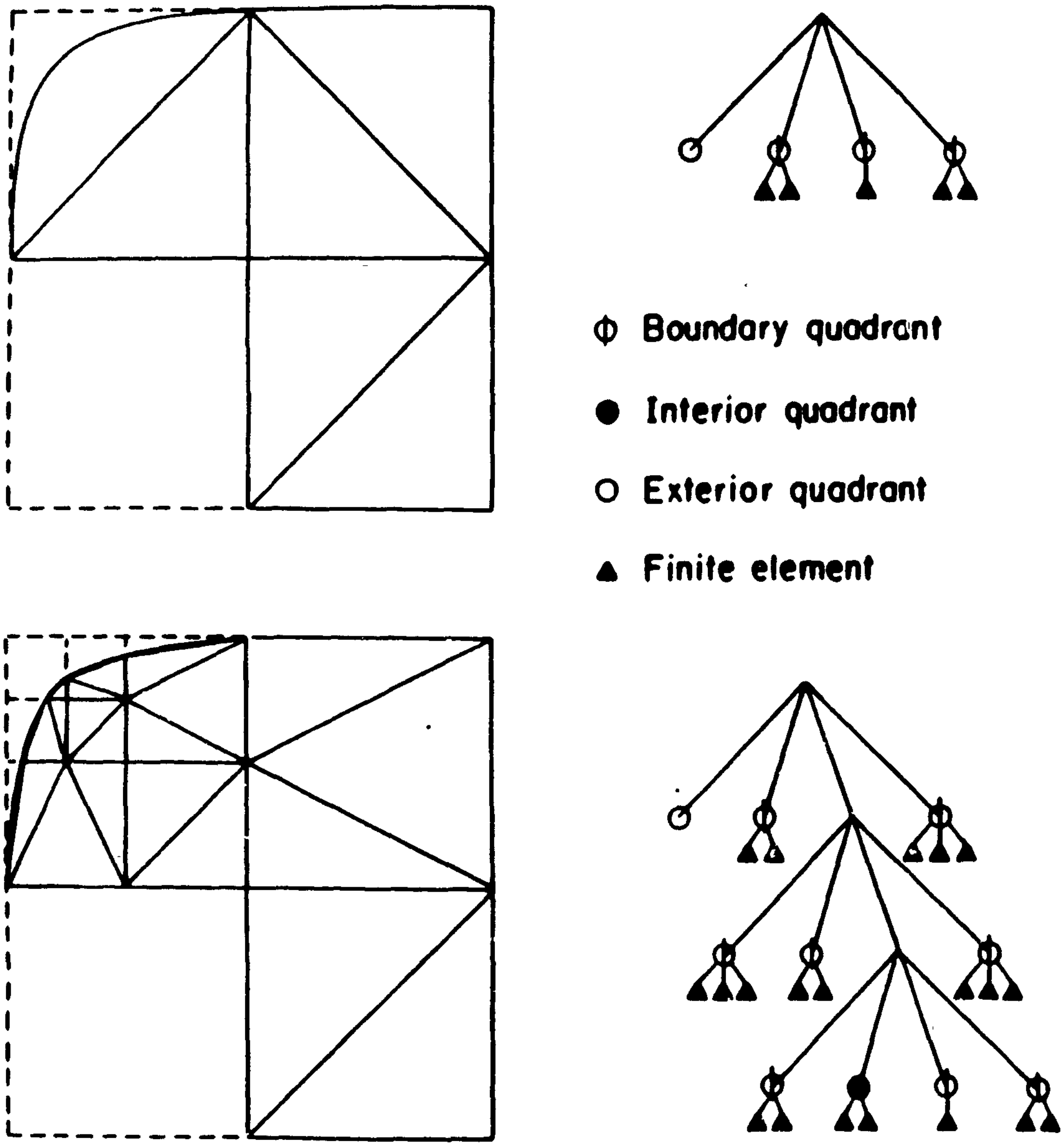


Figure 1. Finite quadtree mesh generation for a domain consisting of a rectangle and a quarter circle. One-level and three-level tree structures and their associated meshes of triangular elements are shown at the top and bottom of the figure, respectively.

phases of the problem under a common tree data structure. Tree depth in such a process will be an explicit function of the prescribed geometric resolution parameters and an implicit function of the refinement indicators present in the adaptive partial differential equations software [5, 6].

2.2. Linear System Solution Strategies

Preconditioned conjugate gradient (PCG) iteration is an efficient means of solving the linear algebraic systems (2.4) that result from the finite element discretization of self-adjoint elliptic partial differential systems [13]. The key steps in the PCG procedure [14] involve (i) matrix-vector multiplication of the form

$$\mathbf{q} = \mathbf{K}\mathbf{p} \tag{2.5a}$$

and (ii) solving linear systems of the form

$$\overline{\mathbf{K}}\mathbf{d} = \mathbf{r}, \tag{2.5b}$$

where $\mathbf{r}$ and $\mathbf{p}$ are the residual vector and conjugate search direction, respectively. The preconditioning matrix $\overline{\mathbf{K}}$ may be selected to reduce computational cost. Reducing the work involved in solving (2.5b), for example, would suggest selecting $\overline{\mathbf{K}}$ close to the identity matrix. This, however, does nothing to reduce the number of conjugate gradient iterations, which would dictate the choice $\overline{\mathbf{K}} = \mathbf{K}$. Naturally, a practical choice of $\overline{\mathbf{K}}$ lies between these two extremes. With our additional goal of developing parallel techniques for solving finite element problems on finite quadtree-structured meshes, we consider (i) an element-by-element (EBE) preconditioning, based on an approximate factorization of $\mathbf{K}$ into the product of elemental matrices, and (ii) a symmetric successive over-relaxation (SSOR) preconditioning.

2.2.1. Element-by-Element Preconditioning

Finite element stiffness matrices $\mathbf{K}$ are summations of elemental contributions; hence, the multiplications in (2.5a) may be performed in parallel in an element-by-element fashion. To be specific, the 3×3 element stiffness matrix $\mathbf{k}_e$ for element e is expanded to the dimension N of S^N as

$$\mathbf{K}_e = \mathbf{C}_e^T \mathbf{k}_e \mathbf{C}_e \tag{2.6a}$$

where $\mathbf{C}_e$ is an $3 \times N$ Boolean connectivity matrix. The matrix-vector product (2.5a), in turn, may be written as

$$\mathbf{q} = (\sum_{e=1}^{N_\Delta} \mathbf{K}_e)\mathbf{p} = \sum_{e=1}^{N_\Delta} \mathbf{C}_e^T \mathbf{k}_e \mathbf{p}_e, \tag{2.6b}$$

where N_Δ is the number of elements in a mesh and $\mathbf{p}_e = \mathbf{C}_e \mathbf{p}$ is the restriction of $\mathbf{p}$ to the unknowns associated with element e. This computation may be done in parallel on noncontiguous elements that have independent basis support; thus, eliminating the need to synchronize critical sections that arise when processes seek simultaneous access to shared data.

Winget and Hughes [15] describe an approximate factorization of the stiffness matrix $\mathbf{K}$ that also relies solely on elemental computations and, hence, would appear to be an appropriate preconditioning to be used in conjunction with the EBE scheduling of the matrix-vector product described above. This factorization has, furthermore, been successfully applied to several finite element computations [16-18]. Carey et al. [19, 20] used other EBE strategies with the PCG method.

To begin, we write $\mathbf{K}$ as

$$\mathbf{K} = \sum_{e=1}^{N_\Delta} \mathbf{K}_e = \mathbf{D}^{½}(\mathbf{I} + \sum_{e=1}^{N_\Delta} \overline{\mathbf{K}}_e)\mathbf{D}^{½} \approx \mathbf{D}^{½}(\mathbf{I} + \varepsilon \sum_{e=1}^{N_\Delta} \overline{\mathbf{K}}_e)\mathbf{D}^{½}, \tag{2.7a}$$

where $\mathbf{D}$ contains the diagonal elements of $\mathbf{K}$, $\overline{\mathbf{K}}_e$ is $\mathbf{K}_e$ less its diagonal elements, and $\varepsilon > 0$ is a parameter to be chosen. Using this representation, the matrix $\mathbf{K}$ is factored approximately as

$$\mathbf{K} \approx \overline{\mathbf{K}} = \mathbf{D}^{½} \prod_{e=N_\Delta}^{1} (\mathbf{I} + ½\varepsilon\overline{\mathbf{K}}_e) \prod_{e=1}^{N_\Delta} (\mathbf{I} + ½\varepsilon\overline{\mathbf{K}}_e)\mathbf{D}^{½}. \tag{2.7b}$$

When $\overline{\mathbf{K}}$ is used as a preconditioner for the conjugate gradient method, solutions of the linear system (2.5b) can be performed in parallel on noncontiguous elements. Thus, the two major computational tasks (2.5a) and (2.5b) present in the PCG can be done in parallel in an element-by-element fashion without synchronizing processes, provided that non-adjacent elements are processed in parallel. Hence, a procedure for computing sets of noncontiguous elements is needed.

2.2.2. SSOR Preconditioning

It is well known [21, 22] that SOR and SSOR iteration can be used for the parallel solution of the five-point finite difference approximation of Poisson's equation on a rectangular mesh by numbering the discrete equations and unknowns in ''red-black'' order. With this ordering, unknowns at red mesh points are only coupled to those at black mesh points and vice versa; thus, solutions at all red points can proceed in parallel followed by a similar solution at all black points. Preserving symmetry, as with SSOR iteration, will make the SOR method a suitable preconditioning for the PCG method.

Adams and Ortega [22] describe multicolor orderings for parallel computation on rectangular grids using finite element and finite difference stencils other than the five-point discrete Laplacian star. However, multicolor orderings for unstructured finite element meshes are more difficult since dependence on nodal connectivity and the polynomial degree of the basis can be quite complex. Finding an ordering to maximize the degree of parallelism would be equivalent to computing the chromatic number of the graph representation of the mesh, which is a known NP-complete problem for arbitrary meshes [23]. Such extreme complexity may be avoided by considering multicolor orderings for block SSOR preconditionings at the quadrant level. To be specific, partition the stiffness matrix **K** by quadrants as

$$\mathbf{K} = \mathbf{D} - \mathbf{L} - \mathbf{L}^T \tag{2.8a}$$

where

$$\mathbf{D} = \begin{bmatrix} \mathbf{K}_{1,1} & & & \\ & \mathbf{K}_{2,2} & & \\ & & \cdots & \\ & & & \mathbf{K}_{\mathcal{Q},\mathcal{Q}} \end{bmatrix}, \quad \mathbf{L} = -\begin{bmatrix} 0 & & & \\ \mathbf{K}_{2,1} & 0 & & \\ & & \cdots & \\ \mathbf{K}_{\mathcal{Q},1} & \mathbf{K}_{\mathcal{Q},2} & & 0 \end{bmatrix}. \tag{2.8b,c}$$

Consider an edge of a triangular element connecting vertices k and l. This edge introduces a nontrivial contribution to block $\mathbf{K}_{i,i}$ of the block diagonal portion $\mathbf{D}$ of the stiffness matrix if nodes k and l are in quadrant i. Contributions to block $\mathbf{K}_{i,j}$ of the lower triangular matrix $\mathbf{L}$ arise when node k lies in quadrant i and node l lies in quadrant j. (The matrices $\mathbf{D}$ introduced in (2.7) and (2.8) have different meanings.)

Using an SSOR preconditioning, the solution of (2.5b) would be computed according to the two-step procedure

$$\mathbf{X}^{n+\frac{1}{2}} = \omega(\mathbf{L}\mathbf{X}^{n+\frac{1}{2}} + \mathbf{L}^T\mathbf{X}^n + \mathbf{r}) + (1-\omega)\mathbf{X}^n, \tag{2.9a}$$

$$\mathbf{X}^{n+1} = \omega(\mathbf{L}^T\mathbf{X}^{n+1} + \mathbf{L}\mathbf{X}^{n+\frac{1}{2}} + \mathbf{r}) + (1-\omega)\mathbf{X}^{n+\frac{1}{2}}, \quad n = 1, 2, \cdots, M. \tag{2.9b}$$

Thus, each block SSOR iteration consists of two block SOR steps; one having the reverse ordering of the other. Typically, $M = 3$ SSOR steps are performed between each PCG step.

Suppose that the Q quadrants of a finite quadtree structure are separated into γ disjoint sets. Then, using the symmetric γ-color block SSOR ordering, we would sweep the quadrants in the order $C_1, C_2, \cdots, C_\gamma, C_\gamma, C_{\gamma-1}, \cdots, C_1$, where C_i is the set of quadrants having color i. Because quadrants rather than nodes are colored, a node can be connected to other nodes having the same color. Thus, the forward and backward SOR sweeps may differ for a color C_i, $i = 1, 2, \cdots, \gamma$. During an SOR sweep, unknowns lying on quadrant boundaries are updated as many times as the number of quadrants containing them.

2.3. Coloring Finite Quadtree Structures

As noted, computation using the PCG method with either the EBE or SSOR preconditioners will be performed in parallel on noncontiguous terminal quadrants of finite quadtree structures. Like the nodal coloring problem referred to in Section 2.2.2, determining noncontiguous elements for arbitrary meshes is equivalent to coloring the vertices of the dual to the planar graph corresponding to the elements of the mesh so that no two vertices have the same color. This is an NP-hard problem [23]. Coloring is greatly simplified by taking advantage of the regular quadtree structure. Thus, a small number of elements associated with each quadrant will have the same color.

Naturally, coloring procedures that use the fewest colors will increase data granularity and reduce the cost of process synchronization. At the same time, the cost of the coloring algorithm should not be the dominant computational cost. With these views in mind, we present eight-color and six-color procedures having linear time complexity. Emphasis will be on the six-color algorithm due to its superior performance.

2.3.1. An Eight-Color Algorithm

Four colors are necessary and sufficient to color a uniform quadtree having all leaf nodes at the same tree level. All that is needed is a simple breadth-first traversal of the tree with assignment of colors numbered 1, 2, 3, 4 to the four leaf nodes having a common parent. Of course, the finite quadtree structure is not generally uniform; however, the one-level difference restriction across quadrant edges (cf. Section 2.1) implies that at most quadrants at three tree levels can intersect at a vertex. Hence, it would be possible to color the tree with twelve colors in three groups of four, e.g., numbered 1 to 4, 5 to 8, and 9 to 12, alternating each set at successive levels of the tree. We consider the possibility of reducing the number of colors to eight by using two sets of four colors alternating through tree levels. Assuming that the orientation of the colors remains the same throughout the process, this strategy fails for the four cases shown in Figure 2 where quadrants having a two-level difference intersect at a common vertex. With the orientation of the colors shown in Figure 2, a simple switch of the colors 5, 6, 7, and 8 to 1, 2, 3, and 4, respectively, at the point where the compromise occurred remedies this difficulty. An example of this successful eight-color procedure is shown in Figure 3.

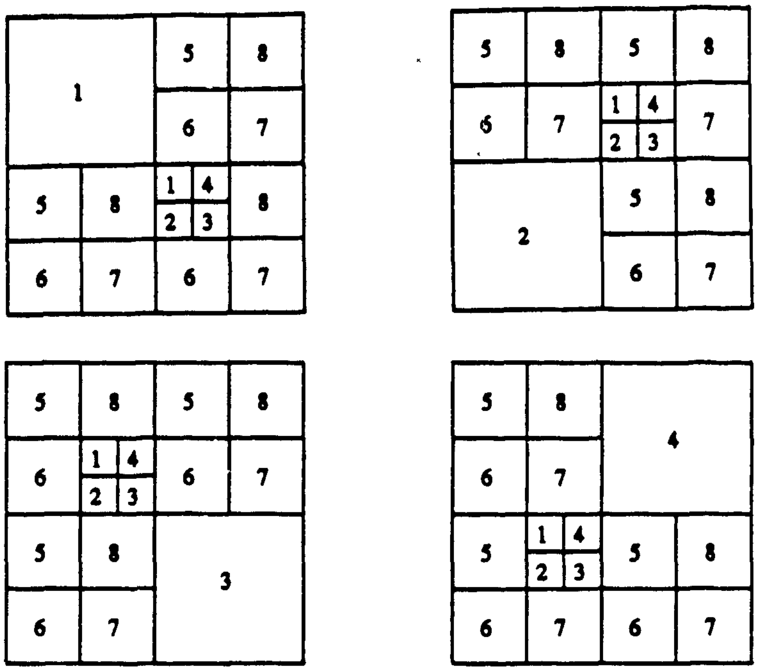


Figure 2. Four possible failures of an eight-color procedure that alternates two groups of four colors through successive tree levels.

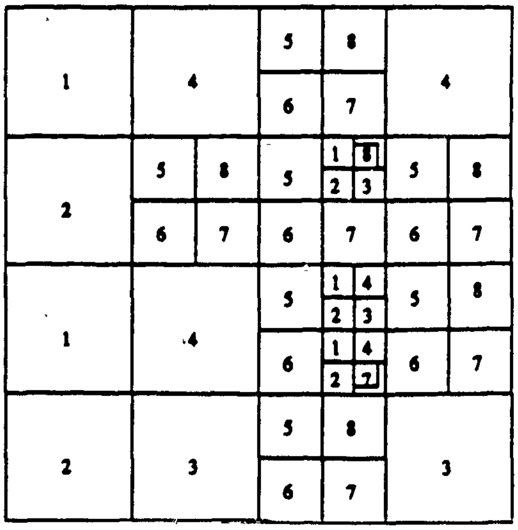


Figure 3. A successful coloring of a quadtree with eight colors.

2.3.2. A Six-Color Algorithm

It is possible to color the quadtree using six colors by performing a "column-order traversal." Towards defining this procedure, let us create a binary directed graph called a "quasi-binary tree" from the finite quadtree by using the following recursive assertive algorithm.

i. The root of the quadtree corresponds to the root of the quasi-binary tree.

ii. Every terminal quadrant is associated with a node in the quasi-binary tree; however, in general, not every quasi-binary tree node corresponds to a quadrant.

iii. In the planar representation of the quadtree, nodes across a common horizontal edge are connected in the quasi-binary tree.

iv. When a quadrant is divided, its parent node in the quasi-binary tree becomes the root of a subtree.

Planar representations of simple quadtrees and their quasi-binary tree representations are illustrated in Figure 4. The leftmost quadtree illustrates root-node and offspring construction of the quasi-binary tree. Connection of nodes across horizontal edges is shown with and without quadrant division in all three illustrations. Subtree definitions according to assertion (iv) are shown in the center and rightmost quadtrees.

From Figure 4 we see that column-order traversal of a finite quadtree is the depth-first traversal of its associated quasi-binary tree. Let us define six colors divided into three sets a, b, and c of two disjoint colors that alternate through the columns in a column-order traversal of the quadtree. Whenever left and right quasi-binary tree branches merge, column-order traversal continues using the color set associated with the left branch. Two of the three color streams, say a and b, are passed to a node of the quasi-binary tree. At each branching, the color stream a and the third color stream c are passed to the left offspring while the streams a and b in reverse order are passed to the right offspring. A

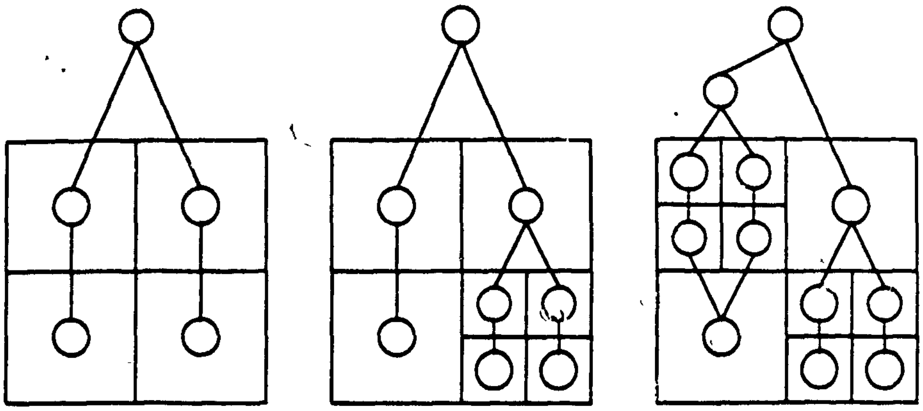


Figure 4. Planar representations of three quadtrees and their associated quasi-binary trees.

```
procedure color_propagate (root: node; $a$, $b$, $c$: color_stream);

  begin
    if not ((root = nil) or (root colored)) then
      begin
        Color root using an alternating color from set $a$;
        color_propagate (left_child, $a$, $c$, $b$);
        color_propagate (right_child, $b$, $a$, $c$)
      end
  end;
```

Figure 5. Color propagation through the quasi-binary tree for the six-color algorithm.

recursive color propagation procedure is described in a pseudo-Pascal language in Figure 5. Assuming that color stream a contains colors 1 and 2, color stream b contains colors 3 and 4, and color stream c contains colors 5 and 6, an example of a planar quadtree colored with the six-color procedure is shown in Figure 6.

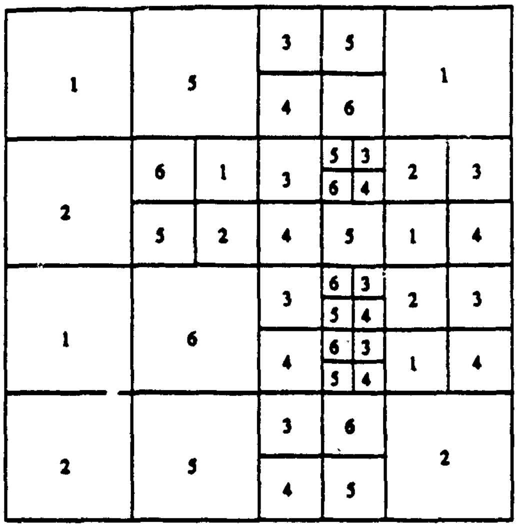


Figure 6. A quadtree colored with the six-color procedure of Figure 5.

2.4. Computational Examples

Solutions of two elliptic systems using PCG iteration with the EBE and SSOR preconditioners and the six- and eight-color algorithms were calculated on a 16-processor Sequent Balance 21000 shared-memory parallel computer. Parallelism is supported through the use of a parallel programming library that permits the creation of parallel processes and enforces synchronization and communication using barriers and hardware locks. Parallel speed up is used as a performance measure.

Example 1. Consider Poisson's equation

$$u_{xx} + u_{yy} = f(x,y), \quad (x,y) \in x^2 + y^2 < 1, \tag{2.10a}$$

with homogeneous Dirichlet boundary conditions applied on the boundary of the unit circle and the function $f(x,y)$ defined such that the exact solution of (2.10a) is

$$u(x,y) = e^{xy} \sin\pi x \sin\pi y. \tag{2.10b}$$

A mesh having 2594 elements, 1346 nodes, and 1099 quadrants was generated. The number of elements per quadrant ranged from 2 to 6. The scalar parameter ε used in the EBE preconditioner was chosen as 0.3 and 3 SSOR iterations were performed per PCG step. Parallel speed up for each preconditioner and coloring scheme is shown in Figure 7. Speed up using the EBE preconditioning is never worse than 73 percent of ideal for the six-color procedure and 70 percent of ideal for the eight-color procedure. Corresponding speed ups for the SSOR preconditioning are 86 and 76 percent of ideal for the six- and eight-color procedures, respectively. Time to calculate solutions of equal accuracy was generally lower with the SSOR preconditioning than with the EBE preconditioning. For example, a solution with 15 processors using the six-color SSOR algorithm used three times less time than the comparable EBE solution. The six-color SSOR procedure also indicates a better scalability to systems having larger numbers of processors.

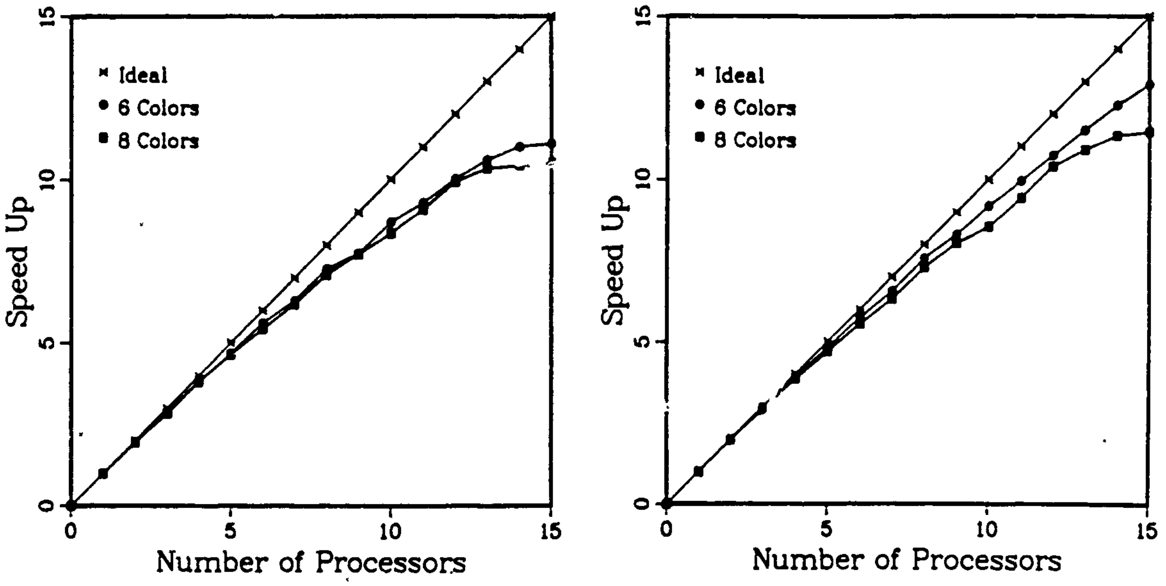


Figure 7. Parallel speed up for Example 1 using the EBE preconditioning (left) and the SSOR preconditioning (right).

Example 2. Consider a plane stress problem satisfying the equations [24]

$$\mathbf{P}^T\mathbf{E}\mathbf{P}\mathbf{u} = \mathbf{f}(x,y), \quad (x,y) \in \Omega, \tag{2.11a}$$

where

$$\mathbf{P} = \begin{bmatrix} \partial/\partial x & 0 \\ 0 & \partial/\partial y \\ \partial/\partial y & \partial/\partial x \end{bmatrix}, \quad \mathbf{E} = \frac{1}{1-\nu^2}\begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & ½(1-\nu^2) \end{bmatrix}. \tag{2.11b}$$

Equation (2.11a) expresses the equilibrium of an elastic continuum having unit Young's modulus and Poisson's ratio ν. The vector $\mathbf{u}(x,y) = [u_1(x,y), u_2(x,y)]^T$ is the displacement of a material point at coordinates (x,y) and $\mathbf{f}(x,y)$ is a body force. The domain Ω and the mesh generated by the finite quadtree procedure are shown in Figure 8. Components of the two-dimensional stress tensor satisfy

$$[\tau_x, \tau_y, \tau_{xy}]^T = \mathbf{EPu}. \tag{2.11c}$$

Boundary conditions are shown symbolically in the upper portion of Figure 8. Thus, the circle, quarter circle and upper boundary are traction free; the stress $Q = 0.5$ on the right boundary; and symmetry conditions apply on the lower and left boundaries. The body force $\mathbf{f} = [0.2, 0]^T$ and $\nu = 0.25$. The mesh shown in the lower portion of Figure 8 has 994 elements, 557 nodes, and 472 quadrants. The EBE parameter ε was chosen as 0.25 and 3 SSOR iterations per PCG step were used.

Parallel speed ups for the EBE and SSOR preconditioners and for each coloring algorithm are shown in Figure 9. Speed up using the EBE preconditioning is greater than 68 percent of ideal for the six-color procedure and 62 percent of ideal for the eight-color procedure. Corresponding speed ups for the SSOR preconditioning are 76 and 71 percent of ideal for the six- and eight-color procedures, respectively.

3. Hyperbolic Problems

Consider the solution of a system of two-dimensional conservation laws in m variables on a rectangular domain having the form

$$\mathbf{u}_t + \mathbf{f}_x(x,y,t,\mathbf{u}) + \mathbf{g}_y(x,y,t,\mathbf{u}) = 0, \quad (x,y) \in \Omega, \quad t > 0, \tag{3.1a}$$

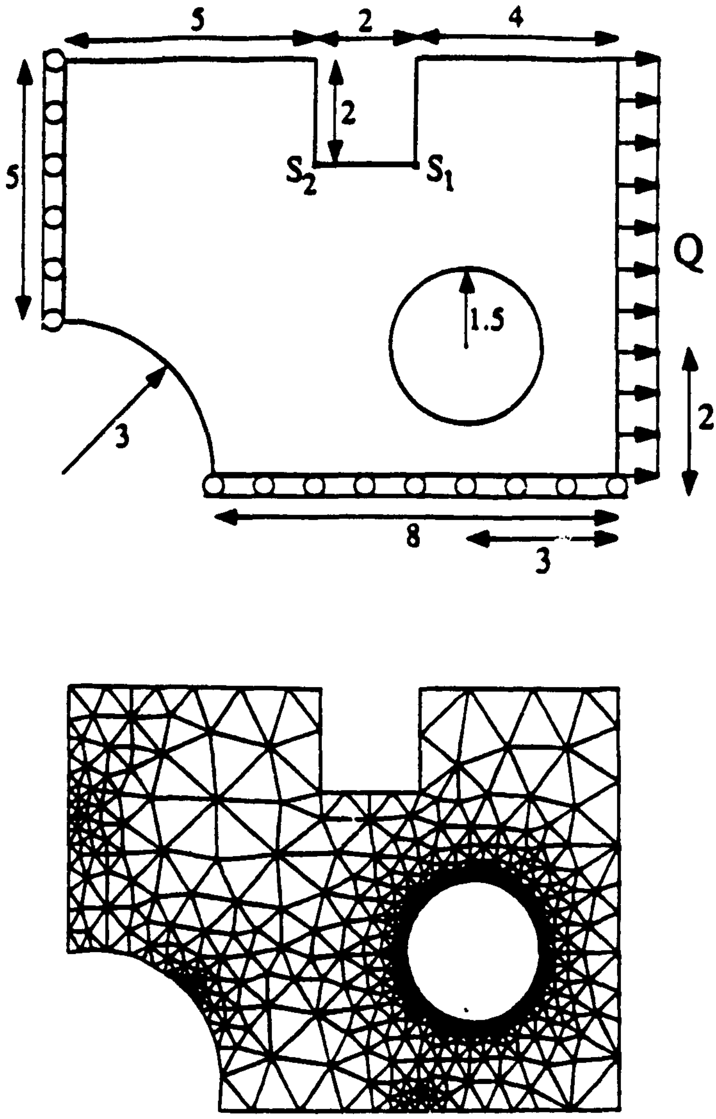


Figure 8. Geometry (top) and mesh (bottom) for Example 2.

subject to the initial conditions

$$\mathbf{u}(x,y,0) = \mathbf{u}^0(x,y), \quad (x,y) \in \Omega \cup \partial\Omega, \tag{3.1b}$$

and appropriate well-posed boundary conditions.

Arney et al. [1, 25] developed an adaptive hr-refinement procedure for solving (3.1) that combined motion of a coarse "base" mesh with recursive local mesh refinement.

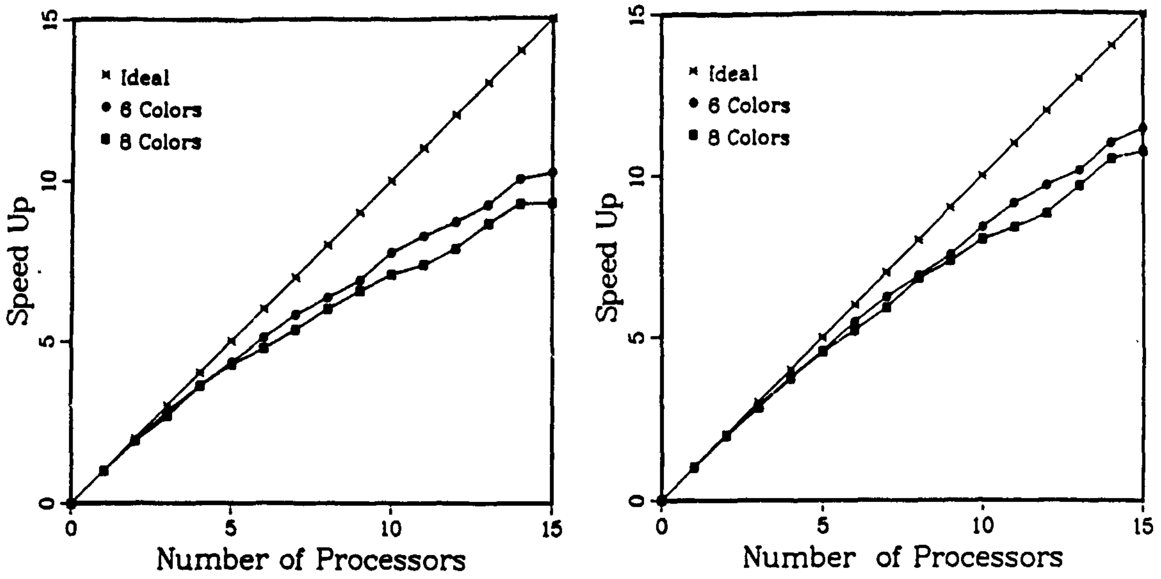


Figure 9. Parallel speed up for Example 2 using the EBE preconditioning (left) and the SSOR preconditioning (right).

Problems were solved on a sequence of base-mesh time slices of duration Δt_n, $n = 0, 1, \cdots$. The discrete solution on a time step from, e.g., t_n to $t_{n+1} = t_n + \Delta t_n$ began with base-mesh motion followed by a finite difference solution and generation of refinement indicators at t_{n+1}. Those cells where the refinement indicator failed to satisfy a prescribed local error tolerance were identified and grouped into rectangular clusters. After ensuring that clusters had an adequate percentage of high-error cells and subsequently enlarging the rectangular clusters by a one-element buffer to provide a transition between high- and low-error regions, cells of the base mesh were bisected in space and time; thus, creating finer meshes associated with each cluster. Problems on the finer meshes were solved and the refinement procedure was repeated until the refinement indicator satisfied the prescribed local error per unit step criteria. All space-time cells of finer meshes were properly nested within those of coarser grids which simplified interpolation problems at mesh interfaces. After finding an acceptable solution on the base mesh, a new base-mesh time step Δt_{n+1} was selected and the integration continued. The mesh motion and local mesh refinement procedures were explicit and independent of each other as well as of the solution technique and motion and refinement indicators. Efficiency suggested the use of a tree structure to manage the data associated with the refinement process. Nodes of the tree corresponded to meshes at each refinement level for the current base-mesh time step. The base mesh was the root node of the tree and finer grids were regarded as offspring of coarser ones.

Arney and Flaherty [1, 4] used MacCormack's finite difference scheme [26] with Davis's artificial viscosity model [27] to obtain discrete solutions on each mesh. This has been replaced by the related Richtmyer two-step version of the Lax-Wendroff method [28], which we describe on a stationary mesh for a one-dimensional problem having no y dependence. Introduce a mesh on Ω having spacing $\Delta x_j = x_{j+1} - x_j$ and let the discrete approximation of $\mathbf{u}(x_j, t_n)$ be denoted as $\mathbf{U}_j^n$. Using the Richtmyer two-step procedure, we predict a solution at the center of the cell $(x_j, x_{j+1}) \times (t_n, t_{n+1})$ using the Lax-Friedrichs scheme

$$\mathbf{U}_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \tfrac{1}{2}(\mathbf{U}_{j+1}^n + \mathbf{U}_j^n) - \frac{\Delta t_n}{2\Delta x_j}(\mathbf{f}_{j+1}^n - \mathbf{f}_j^n). \tag{3.2a}$$

This provisional solution is corrected using the leap frog scheme

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{2\Delta t_n}{\Delta x_j + \Delta x_{j-1}}(\mathbf{f}_{j+\frac{1}{2}}^{n+\frac{1}{2}} - \mathbf{f}_{j-\frac{1}{2}}^{n+\frac{1}{2}}). \tag{3.2b}$$

The motion and refinement schemes are not limited to either the MacCormack or the Richtmyer two-step methods and their selection was based on a desire for generality rather than for optimal performance on a specific application.

Arney et al. [1, 25] used estimates of the local discretization error, obtained by Richardson's extrapolation on a space-time mesh having half the spacing of the current mesh, as a refinement indicator. Solutions generated on the finer mesh as part of the error estimation process could subsequently be used as a fine-mesh solution when refinement was necessary. Initial and boundary data for refined meshes was determined by bilinear interpolation from acceptable solutions on the finest available meshes.

3.1. Parallel Solution Strategies

Because of the global effects introduced by the migration of computational cells, we postpone a treatment of mesh motion and herein appraise the suitability of Arney et al.'s [1, 25] h-refinement scheme for parallel computation on a p-processor shared-memory MIMD computer. Two possible parallel computation strategies immediately come to mind: (i) depth-first traversal of the tree of grids with each grid being processed in parallel and (ii) parallel processing of solutions on grids at the same tree level. Using depth-first traversal, a static domain decomposition of each grid into p subregions is performed and a processor is assigned to each subregion. With the second alternative, the p processors are distributed among the grids at a level. The processors that are assigned to a grid are released and reassigned elsewhere when refinement of a particular grid terminates. Thus, with this strategy, parallelism occurs both within a grid and across the breadth of the tree. In either case, the parallel solution process proceeds from one base-mesh time step to the next.

Serial depth-first traversal of the tree leads to a highly structured algorithm that has a straight-forward design because the same procedure is applied to all grids. Balancing processor loading on rectangular grids would be nearly perfect with an explicit finite difference scheme such as (3.2). Balancing loads on geometrically complex regions would be more difficult, but still would only require a static decomposition of a grid into p subdomains each having nearly the same number of computational cells. Load imbalance occurs even with rectangular grids due to the differences in the time required to compute initial data. Other than at $t = 0$, initial data is determined by interpolating solutions on the finest grid at the end of the previous base-mesh time step. Traversal of the tree structure is needed to determine the correct solution points for the interpolation. Such a traversal could take different times in different regions due to v: ations in tree depth. Additionally, the interpolation is more or less complex depending on whether solution points do not or do coincide. We will offer additional comments on this issue in Section 3.3.

The serial depth-first traversal procedure would become inefficient when p exceeds the number of elements in a grid. This possibility can be reduced by refining by more than a binary factor; thus, maintaining a shallow tree of finer grids. Lowering the efficiency of clusters, thereby including a greater percentage of low-error cells, will also increase grid size but diminish optimal grid usage. With resolution, we would expect this serial traversal procedure to be also viable on data-parallel computers.

The parallel tree traversal procedure requires complex dynamic scheduling procedures to assign processors to grids. As discussed in Section 3.3, this may potentially be done by estimating the work remaining to reduce errors to prescribed tolerances on all subgrids at a given tree level and assigning processors accordingly. Were such a heuristic load balancing technique successful, we would not expect the parallel tree traversal procedure to degrade in efficiency when the number of elements on a grid is $O(p)$.

3.2. A Computational Example

Let us exhibit the results of a computational experiment applying the serial depth-first procedure to an example. Simplifying assumptions listed below have been incorporated into the algorithm.

i. Each grid is statically decomposed into p subgrids having nearly the same number of computational cells. A one-element overlap is included at the boundaries between adjacent pairs of subdomains so that processors can proceed asynchronously within their respective subdomains. No attempt is made to balance processor loading based on the complexity of the initial data at the beginning of a base-mesh time step.

ii. Unlike the serial procedure [1], all grids are rectangular with edges parallel to the coordinate axes.

iii. Grids spawned by refinement must remain within the boundaries of their parent subgrid.

A computer code based on this algorithm has been implemented on the 16-processor Sequent Balance 21000 computer and applied to the following problem.

Example 3. Consider the linear scalar hyperbolic differential equation

$$u_t + u_x + \frac{1}{4}u_y = 0, \quad 0.2 \le x \le 1.2, \quad 0 \le y \le 1, \quad t > 0, \tag{3.3a}$$

with initial and Dirichlet boundary data specified so that the exact solution is

$$u(x,y,t) = \begin{cases} 0.8, & \text{if } (y-0.25\,t) < -4(x-t) + 1.2 \\ 0, & \text{if } (y-0.25\,t) > -4(x-t) + 1.6 \\ -8(x-t) - 2(y-0.25t) + 3.2, & \text{otherwise.} \end{cases} \tag{3.3b}$$

Equation (3.3b) is an oblique ramp-like wave front that moves at an angle of 14 degrees across the domain as time progresses.

Refinement was controlled by using an approximation of the local discretization error in the L^1 norm as a refinement indicator. Exact errors for this scalar problem were also measured in L^1 as

$$\|e(\cdot,\cdot,t)\|_1 = \iint_\Omega |Qu(x,y,t) - U(x,y,t)|\, dxdy, \tag{3.4}$$

where $U(x,y,t)$ is a piecewise constant representation of the discrete solution and $Qu(x,y,t)$ is a projection onto the space of piecewise constant functions obtained by using values at cell centers.

We solved this problem on $0 \le t \le 4.2$ using local refinement restricted to 0, 1, and 2 tree levels. The term "0 refinement levels" implies no adaptivity. A 20×20 base mesh, an initial time step of 0.032, and a refinement tolerance of 0.002 was used in all cases. Parallel speed up for each strategy is shown in the left portion of Figure 10. As an indication of the maximum speed up possible on a Balance 21000 computer, we also solved a problem with no load imbalance due to initial conditions on a 6×360360 grid. Thus, the only degradation from ideal speed up would be due to bus contention. Speed up for this "embarrassingly parallel" problem is shown in the right portion of Figure 10. These results indicate that it is possible to obtain approximately 90 percent of ideal speed up for our procedure on a Balance 21000 computer. We would like to regard this as the upper limit to the parallel performance of the current version of our adaptive algorithm.

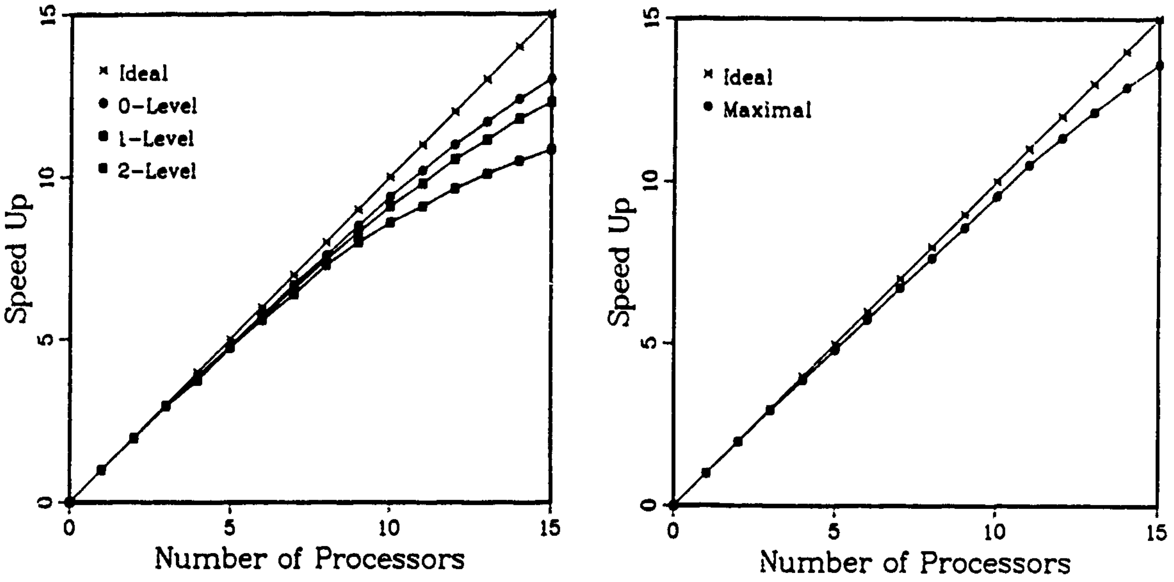


Figure 10. Parallel speed up for Example 3 using refinement restricted to 0, 1, and 2 tree levels (left, top to bottom) and for a perfectly balanced problem (right).

Maximum speed ups shown in Figure 10 are greater than 86, 82, and 72 percent of ideal for problems having 0, 1, and 2 levels of refinement, respectively. Speed ups relative to the maximum feasible speed up reported on the right of Figure 10 are, respectively, 96, 90, and 79 percent for 0, 1, and 2 refinement levels. Our adaptive procedures are capable of achieving a high degree of parallelism; however, performance degrades as tree depth increases due to the serial overhead incurred when managing a more complex data structure.

Speed up is not an appropriate measure of the complexity required to solve a problem to a prescribed level of accuracy. Tradeoffs occur between the higher degree of parallelism possible with a uniform mesh solution and the greater sequential efficiency of an adaptive procedure. In order to gage the differential, we computed uniform mesh and adaptive mesh solutions of Example 3 on various processor configurations and to varying levels of accuracy. Computations on uniform meshes ranged from a 10×10 mesh to a 90×90 mesh. All adaptive computations used a 20×20 base mesh and an unrestricted number of refinement levels.

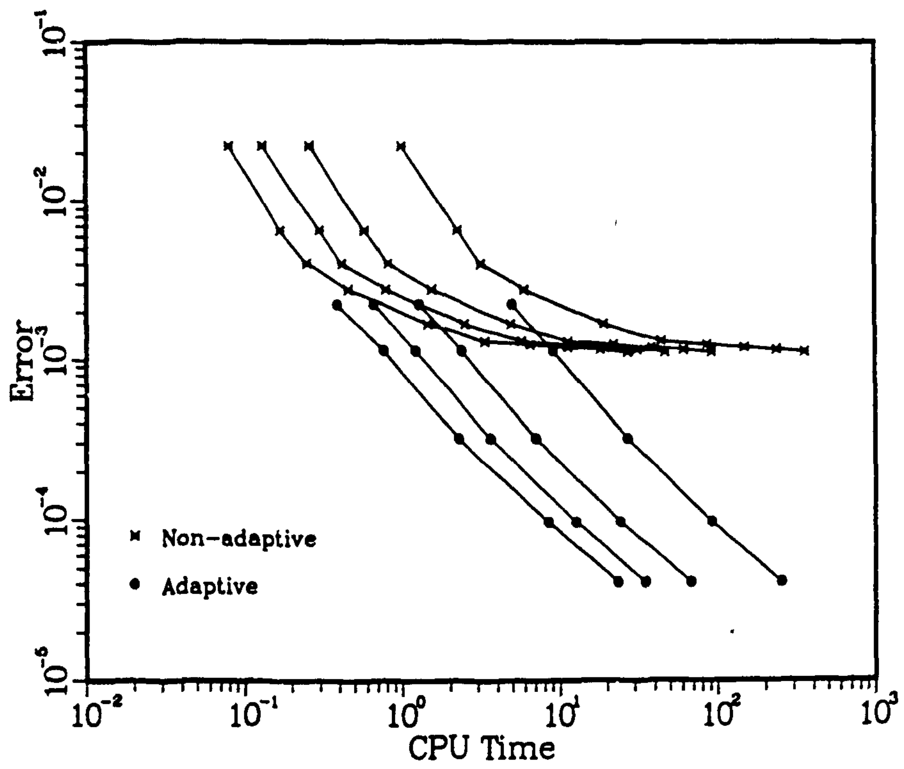


Figure 11. Global L^1 error as a function of CPU time for Example 3 using non-adaptive methods (upper set of curves) and adaptive h-refinement methods (lower set of curves). Each computation was repeated using 1, 4, 8, and 15 processors (right to left in each set of curves).

Results for the global L^1 error as a function of effort (CPU time) are presented in Figure 11 for computations performed on 1, 4, 8, and 15 processor systems. The upper set of curves, displaying non-adaptive results, are much less efficient than the adaptive solutions shown on the lower portion of the figure. Bus saturation due to the large volume of data has limited non-adaptive solutions to unacceptably low levels of accuracy. Solutions obtained on systems having greater bus bandwidth would postpone this problem and diminish the advantages of an adaptive approach; however, the difficulty would still arise with increasing problem complexity.

3.3. Parallel Tree Traversal

We conclude this section with a brief discussion of parallel tree traversal procedures. Consider a situation where q processors were used to obtain a solution on a grid at tree level $l-1$ and suppose that refinement indicators dictate the creation of L level l grids. Further assume that (i) the prescribed local refinement tolerance at level l is τ; (ii) the areas of the level l grids $G_{l,i}$ are $M_{l,i}$, $i = 1, 2, \cdots, L$; (iii) error estimates $E_{l,i}$, $i = 1, 2, \cdots, L$, can be obtained for each grid from the level $l-1$ refinement indicators; and (iv) the finite difference solution is converging as the square of the local mesh spacing. Quadratic convergence is established merely as an example and the approach easily extends to other convergence rates.

In order to satisfy the prescribed accuracy criterion, the spatial domain $G_{l,i}$ should be refined by a factor of $(E_{l,i}/\tau)^2$. The time step on $G_{l,i}$ must also be reduced by a factor of $E_{l,i}/\tau$ in order to satisfy the Courant condition. Hence, the expected work $W_{l,i}$ to find an acceptable solution on the region $G_{l,i}$ is

$$W_{l,i} = M_{l,i}(\frac{E_{l,i}}{\tau})^3. \tag{3.5}$$

The original q processors should be allocated so as to balance the time required to complete the expected work on each of the L grids at level l. Thus, q_i, $i = 1, 2, \cdots, L$, processors should be assigned to the level l grids so that

$$\frac{W_{l,1}}{q_1} = \frac{W_{l,2}}{q_2} = \cdots = \frac{W_{l,L}}{q_L}, \quad \sum_{i=1}^{L} q_i = q. \tag{3.6a,b}$$

Quality of load balancing by this approach will depend on the accuracy and robustness of the error estimate. Previous investigations [1, 25] revealed that error estimates were generally better than 80 percent of the actual error for a wide range of mesh spacings and problems. Equation (3.5) can be used to select refinement factors other than binary and, indeed, to select different refinement levels for different meshes at a given tree level. This consideration combined with over-refinement to a tolerance somewhat less than the prescribed tolerance should maintain a shallow tree depth and enhance parallelism at the expense of grid optimality.

The procedures outlined by (3.5) and (3.6) can additionally be modified to balance nonuniformities in the initial data at the beginning of base-mesh time steps for $t > 0$. Thus, the work estimate (3.5) could be multiplied by a factor representing the work needed to calculate initial data. The additional effort required to acquire the initial data, assuming that such access is as complex as computing the solution, increases the expected work during the first of $E_{l,i}/\tau$ fine time steps by $(E_{l,i}/\tau)^2$. Hence, the revised estimated work is

$$\overline{W}_{l,i} = W_{l,i}[1 + \frac{E_{l,i}}{\tau} - \frac{\tau}{E_{l,i}}]. \tag{3.7}$$

4. Discussion

We have described parallel finite element procedures for solving elliptic problems on finite quadtree structured grids and parallel adaptive procedures for the explicit finite difference solution of hyperbolic problems. Speed ups on a shared-memory computer are used to demonstrate that a high degree of parallelism has been established in all cases. Decay in speed up after using approximately 8 of the 15 available processors on a 16-processor Sequent Balance 21000 computer occurs due, for example, to loss of processor synchronization, start-up latency, and large data granularity.

For elliptic problems, a six-color scheme for separating elements at the quadrant level of a finite quadtree structure has better speed up and scalability than an eight-color scheme, which is expected due to the decreased need for process synchronization. The number of quadrants within a color group may not be divisible by the number of processors, which may result in some processors being idle near the end of each task queue. Similar losses of efficiency could result due to differences in the number of elements per quadrant. Although the six-color procedure requires less communication, these reasons may account for the small difference in speed up between the six- and eight-color schemes on Example 2.

Examples 1 and 2 indicate a higher speed up for the SSOR preconditioning than for the EBE preconditioning. This may be due to a need to use some global information in the preconditioner [29]. Further improvements of the EBE preconditioning are likely by adjusting the parameter ε in (2.7). In any event, performance of the EBE preconditioner with $\varepsilon > 0$ is better than the diagonal scaling preconditioner that results when $\varepsilon = 0$.

Parallel speed up of our adaptive h-refinement scheme for hyperbolic systems degrades as tree depth increases. Nevertheless, adaptive tree data structures utilize less data than uniform structures which lowers contention on a bus-based multi-processor. This enabled solutions of Example 3 to be calculated to much greater accuracy than with uniform structures having a higher degree of parallelism.

Our schemes for both elliptic and hyperbolic systems are far from being complete and several computational and theoretical issues are yet to be resolved. Currently, hierarchical bases and p- and hp-refinement techniques are being added to both systems. Hierarchical bases are, of course, well established for elliptic systems [3]. Their use for hyperbolic problems can be done by an approach of Cockburn and Shu [30]. The EBE preconditioning should continue to exhibit a high degree of parallelism with the higher-order bases; however, load balancing of adaptive hp-refinement schemes will present quite a challenge.

The present Richardson's extrapolation-based error estimation technique used to furnish local error estimates of solutions to hyperbolic systems is expensive and will be replaced by a technique based on p-refinement. Computation at the beginning of each base-mesh time step need not return to the base mesh, but could begin on an adaptively chosen mesh that utilizes known nonuniformities in the solution discovered during the previous base-mesh time step. Scheduling processors to balance loads in this case is also far from clear.

Finally, both adaptive procedures for elliptic and hyperbolic systems will be attempted on distributed memory computers.

References

[1} D.C. Arney and J.E. Flaherty, An adaptive mesh-moving and local refinement method for time-dependent partial differential equations, ACM Trans. Math. Softw., to appear.

[2] R.E. Bank, PLTMG Users' Guide, Tech. Rep., Department of Mathematics, University of California, San Diego, 1981.

[3] I. Babuska, B.A. Szabo, and I.N. Katz, The p-version of the finite element method, SIAM J. Numer. Anal. 18 (1981) 515-545.

[4] D.C. Arney and J.E. Flaherty, A two-dimensional mesh moving technique for time-dependent partial differential equations, J. Comput. Phys. 67 (1986) 124-144.

[5] S. Adjerid and J.E. Flaherty, A moving-mesh finite element method with local refinement for parabolic partial differential equations, Comput. Meths. Appl. Mech. Engrg. 55 (1986) 3-26.

[6} S. Adjerid, J.E. Flaherty, and Y. Wang, A posteriori error estimation and adaptive refinement with finite element methods of lines for one-dimensional parabolic systems, in preparation.

[7] I. Babuska and T. Janik, The h-p version of the finite element method for parabolic equations. Part I - The p-version in time, Rep. MD88-20-IB-TJ, TR88-20, Institute for Physical Science and Technology, University of Maryland, College Park, 1988.

[8] I. Babuska and T. Janik, The h-p version of the finite element method for parabolic equations. Part II - The h-p version in time, Rep. MD89-04-IB-TJ, TR89-04, Institute for Physical Science and Technology, University of Maryland, College Park, 1989.

[9] J. Devloo, J.T. Oden, and P. Pattani, An h-p adaptive finite element method for the numerical simulation of compressible flow, Comput. Meths. Appl. Mech. Engrg. 70 (1988) 203-235.

[10] I. Babuska and E. Rank, An expert-system like approach in the hp-version of the finite element method, Tech. Note BN-1084, Institute for Physical Science and Technology, University of Maryland, College Park, 1986.

[11] R.E. Bank and A. Weiser, Some a posteriori error estimators for elliptic partial differential equations, Math. Comp. 44 (1985) 283-301.

[12] P.L. Baehmann, S.L. Wittchen, M.S. Shephard, K.R. Grice, and M.A. Yerry, Robust, geometrically based, automatic two-dimensional mesh generation, Internat. J. Numer. Meths. Engrg. 24 (1987) 1043-1078.

[13] O. Axelsson and V.A. Barker, Finite Element Solution of Boundary Value Problems: Theory and Computation (Academic Press, Orlando, 1984).

[14] J.M. Ortega, Introduction to Parallel and Vector Solution of Linear Systems (Plenum Press, New York, 1988).

[15] J.M. Winget and T.J.R. Hughes, Solution algorithms for nonlinear transient heat conduction analysis employing element-by-element iterative strategies, Comput. Meths. Appl. Mech. Engrg. 52 (1985) 711-815.

[16] I. Gustafsson and G. Lindskog, A preconditioning technique based on element matrix factorizations, Comput. Meths. Appl. Mech. Engrg. 55 (1986) 201-220.

[17] R.B. King and V. Sonnad, Implementation of an element-by-element solution algorithm for the finite element method on a coarse-grained parallel computer, Comput. Meths. Appl. Mech. Engrg. 65 (1987) 47-59.

[18] B. Nour-Omid and B.N. Parlett, Element preconditioning using splitting techniques, SIAM J. Sci. Stat. Comput. 6 (1985) 761-770.

[19] G.F. Carey and B.-N. Jiang, Element-by-element linear and nonlinear solution schemes, Comm. Appl. Numer. Meths. 2 (1986) 145-153.

[20] G.F. Carey, E. Barragy, R. McLay, and M. Sharma, Element-by-element vector and parallel computations, Comm. Appl. Numer. Meths. 4 (1988) 299-307.

[21] L. Hayes, Comparative analysis of iterative techniques for solving Laplace's equation on the unit square on a parallel processor, M.S. Thesis, University of Texas, Austin.

[22] L. Adams and J. Ortega, A multi-color SOR method for parallel computation, in: K.E. Batcher, W.C. Meilander, and J.L. Potter, eds., Proceedings of the International Conference on Parallel Processing (Computer Society Press, Silver Spring, 1982) 53-56.

[23] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness (Freeman, San Francisco, 1979).

[24] E.B. Becker, G.F. Carey, and J.T. Oden, Finite Elements: An Introduction (Prentice-Hall, Englewood Cliffs, 1981).

[25] D.C. Arney, R. Biswas, and J. E. Flaherty, An adaptive mesh moving and refinement procedure for one-dimensional conservation laws, in preparation.

[26] R.W. MacCormack, The effect of viscosity in hypervelocity impact cratering, AIAA Paper No. 69-354, 1969.

[27] S.F. Davis, A simplified TVD finite difference scheme via artificial viscosity, SIAM J. Sci. Stat. Comput. 8 (1987) 1-18.

[28] R.D. Richtmyer and K.W. Morton, Difference Methods for Initial-Value Problems (Interscience, New York, 1967).

[29] D.E. Keyes and W.D. Gropp, A comparison of domain decomposition techniques for elliptic partial differential equations and their parallel implementation, in: C.W. Gear and R.G Voigt, eds., Selected Papers from the Second Conference on Parallel Processing for Scientific Computing (SIAM, Philadelphia, 1987) s166-s202.

[30] B. Cockburn and C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework, Math. Comp. 52 (1989) 411-435.

The Three R's of Engineering Analysis

and

Error Estimation and Adaptivity

O.C.Zienkiewicz and J.Z.Zhu

Institute of Numerical Methods in Engineering, University College of Swansea, SA2 8PP, U.K.

Abstract

Reason, Robustness and reliability must be the necessary guidelines for practical computer based engineering analysis. With the increasing use of 'black box' computer codes incorporating finite element approximations it is essential that at least the second two attributes be given considerable attention by code designers. Here the role of error estimation and automatic adaptivity is of particular importance and the paper reviews the state of the art and possibilities now available.

1. The three R's

The widespread availability of computational tools, based on finite elements and other forms of approximation, opens almost unlimited possibilities for the progress of engineering. Previously insoluble problems can today be treated on a routine basis and, at least in principle, this should lead to immense improvement in design. However, many dangers are apparent as, with the 'black box' approach, facilities are placed within reach of a wide community often lacking the necessary background knowledge.

Just as the use of 'sharp tools' presents a danger to children, so powerful codes may prove disastrous when applied without suitable education and safeguards.

In preparing our children to face the world their education is based on the well known three R's foundations of Reading, wRiting and aRithmetic. In a similar way we believe that the development and use of analysis codes should be guided by another set of R's -- i.e Reason, Robustness and Reliability. The first should be in the main addressed to the user. The remaining are firmly in the hands of code developers.

The conference at which this paper was initially presented addressed all of the above but in the present context we feel it is of importance to clarify the terms.

1.1 Reason, or alternatively knowledge and intelligence is, without doubt, the most important of the three items. It must guide the decisions concerning the validity of mathematical method used in simulating the physical problem of interest and those of assessing the results achieved. Here, despite many developments of 'knowledge based systems' and 'artificial intelligence' the professional education and wisdom of the engineer can not be superseded. This indeed is fortunate as otherwise his role could almost be eliminated after the formulation of a conceptual design.

The questions which arise here and which must be answered without ambiguity are numerous. The objectives of the analysis and the nature of answers sought must first be clearly visualized. Sometimes these may be found in appropriate codes of practice of the profession -- but more frequently their ambiguity makes it necessary for the engineers to define and justify his own aims and this definition is not trivial. With objectives defined the process of modeling begins with such questions as: Is this a problem of solid or fluid mechanics? Will a linear elastic solution be sufficient or do plastic effects intervene? etc. Again the answers are not at all obvious. Such problems as those posed by metal forming where deformations are very large may be modeled either as solids or fluids. Indeed the latter approach is frequently advantageous.

Even when linear elasticity is chosen as a model many pitfalls are present if the knowledge of the background is not possessed by the user. Professor Babuska and Bathe have, elsewhere in this issue shown many examples of difficulties encountered in deciding how the elastic model is to be used. Fig.1 to 3 show some further situation in which difficulties may occur to the 'uninitiated'.

(FIGURE 1)

In the first (Fig. 1) the problem is that of determining the foundation displacements under a weight at a building. As the building is long a two dimensional plane strain idealization is considered and, as finite elements are to be used in the analysis the user experiments with different depths of the foundation included in the analysis. The fact that the actual displacements will be infinite in the idealization chosen escapes his attention! (Of course the problem is not meaningless if only relative displacements between different parts of the foundation are of interest)

(FIGURE 2)

In Fig.2 a plate bending problem is encountered; here depending on the relative thickness/span ratio either thin (Kirchhoff) or thick (Reissner-Mindlin) plate theories may be relevant. It is now well established that the former can give very different results from the latter in what usually have been considered thin cases when so called 'simple support' conditions are specified [1][2]. However the simple support found in practice (and shown in the figure) differs much from that assumed in plate theory.

The differences in a thick plate situation (Fig. 2a) are such that three dimensional effects can add considerably to deformations and cause stresses not existing in plate theory. Further the determination of plate 'spans' is by no means precise. Although these effects may be insignificant when very thin plates are analyzed for these unless deflections are very small, membrane effects will again mask predictions of linear plate theory. The user has to be sufficiently educated to decide on the limitations.

(FIGURE 3)

The last example Fig.3 shows a problem of a shaft subject to torsion which could (wrongly) be used to prove non uniqueness of the elastic solutions. Here a rounded keyway or notch is examined and stresses computed with a very precise finite element analysis. Irrespective of the size of the notch the analysis will show almost identical maximum stresses and these will persist as the size of the notch tends to zero. These stresses are very different from those found in the same shaft with no notch! Are the final results nonunique? Microscopic examination of any surface shows up many imperfections are therefore notched results may be more realistic!

Indeed if a sharp notch is considered stresses near its tip will always be infinite! Are infinite stresses therefore always present in the limit?

The answer here lies, we believe, in 'size' limitations beyond which continuum theories cease to apply and this is tacitly assumed by skilled engineers. Can we guarantee that the user of codes has the necessary skills?

Clearly in most of the problems of modeling and interpretation considerable knowledge is required of the user and computer knowledge bases can only provide minimum guidance. Fortunately the case of the remaining two R's is more likely yield to the design of computer codes and here we expect greater strides in the future.

1.2 Robustness is a word which has recently been introduced into the numerical analysis vocabulary without perhaps a proper definition. In the present context it is simply the requirement that computational algorithm used in codes should work in all cases falling within their class of applicability. Examples of lack of robustness abound in finite element literature with elements purporting to solve a particular problem, and shown to do so in n - demonstration cases, failing in the n+1 application (which was not expressly excluded in their design). Here we shall not quote examples which could prove embarrassing but shall state simply that robustness can be assured by the code developers by subjecting the elements to a properly conceived patch test [3-5]. The same can alternatively be assured, albeit with more difficulty, if at the stage of element conception the mathematical reasoning of the Babuska-Brezzi test is applied [6,7].

Though lack of robustness may not be detected at the early stages of algorithm development by its well intentioned originators, the well know McPherson law will ensure that, if at all possible, it will fail at the most inconvenient moment.

Clearly robustness of all algorithms should be assured as far as possible by the codes so that the user, who justifiably need not be an expert in numerical analysis, should not be disappointed.

1.3 Reliability is the last of the three R's and most of this paper will be devoted to this aspect. With robustness already taken care of the remaining causes of results on which the user can not rely are (i) programming bugs (ii) errors due to roundoff

(iii) errors of discretization. The first two though impossible to eliminate entirely are well recognized and for many years have been tackled by code designers introducing appropriate checks. However the discretization error, which is always present in numerical approximation process, has been considered beyond the scope of standard coding and its control left to the user. Here an obvious difficulty is encountered by the unskilled user with lack of numerical analysis knowledge - but fortunately now the situation can be remedied by the procedure of Error Estimation informing the user of errors present in the analysis executed and Adaptive Refinement giving, hopefully automatically, results with user specified accuracy. In the following sections we shall discuss the state of the art of these in some detail.

2. Error Estimation and Adaptivity -- What is now available!

2.1 The error and its assessment

The errors in any numerical discretization process can only be assessed 'a posteriori' i.e after a numerical solution has been carried out (though some qualitative guidance regarding, say, the mesh size to be adopted can be gained from so called 'a priori' estimates) Much research has been devoted to the subject since the mid seventies through the pioneering work of Babuska and others [8-20]. Recently error estimating procedures have been made considerably easier at least for linear elliptic problems through the work of Zienkiewicz, Zhu and others [21-31,2]. The state of the art today is such that for many problems of practical interest the estimate of errors can be made after the completion of the solution at a fraction of its cost.

We are somewhat surprised that it has taken a very considerable time for code developers to incorporate such procedures which would at least shown the user the accuracy he has achieved and the reliability of the answers.

Indeed it is now possible to devise a post processor which can take results of an analysis carried out on any existing code and supply the error information. Such code called TAP-2D (Test And Predict) is now commercially available.

We shall illustrate the Z^2 (Zienkiewicz and Zhu) error estimating and adaptive mesh design processes on a general elliptic problem of the type

$$\underline{S}^T \underline{D}\,\underline{S}\,\underline{u} = \underline{f} \tag{1}$$

with appropriate boundary conditions incorporated.

Here $\underline{S}$ is a general differential operator which in the particular case of elasticity defines strains as

$$\underline{\varepsilon} = \underline{S}\,\underline{u} \tag{2a}$$

and stresses as

$$\underline{\sigma} = \underline{D}\,\underline{S}\,\underline{u} \tag{2b}$$

If a finite element solution using standard procedures [32] is obtained with a trial function expansion of the form

$$\underline{u}^h = \underline{\underline{N}}\bar{\underline{u}} \qquad (3)$$

then local error is

$$\underline{e} = \underline{u} - \underline{u}^h \qquad (4)$$

This has to be specified in some norm for simple control purposes and generally the so called energy norm is used. Here we define this as

$$\|\underline{e}\| = \left(\int_\Omega (\underline{\underline{S}}\underline{e})^T \underline{D} (\underline{\underline{S}}\underline{e}) \, d\Omega \right)^{1/2} \qquad (5a)$$

or perhaps more understandably as

$$\|\underline{e}\| = \left(\int_\Omega \underline{e}_\sigma^T \underline{D}^{-1} \underline{e}_\sigma \, d\Omega \right)^{1/2} \qquad (5b)$$

where

$$\underline{e}_\sigma = \underline{\underline{DS}}\underline{u} - \underline{\underline{DS}}\underline{u}^h = \underline{\sigma} - \underline{\sigma}^h \qquad (5c)$$

is the error in stresses for an elastic problem

> Remark In other elliptic problems of frequent usage precisely the same norm can be, of course, applied. For instance in a heat conduction problem the variable change

$\underline{S} \longrightarrow \underline{\nabla}$, $\underline{D} \longrightarrow k$ (conductivity)

$\underline{u} \longrightarrow T$, (temperature) and $\underline{\sigma} \longrightarrow \underline{q}$ (heat flux)

will accomplish the modifications needed.

The evaluation of error in energy norm by above expression is still of course impossible as knowledge of an exact solution is lacking. However if by some procedure values of, say, $\underline{\sigma}^*$ which are approximately an order of h more accurate than those of $\underline{\sigma}^h$ can be found, then the integrals of equation (5b) could be determined by putting

$$\underline{\sigma} \cong \underline{\sigma}^* \qquad (6)$$

and an estimate of error obtained.

In the procedure discussed here such approximation are obtained by using a smooth representation of $\underline{\sigma}^*$

$$\underline{\sigma}^* = \underline{N}\bar{\underline{\sigma}}^* \qquad (7)$$

and ensuring that the projections of $\underline{\sigma}^*$ and $\underline{\sigma}^h$ coincide i.e that

$$\int_\Omega \underline{P}^T(\underline{\sigma}^* - \underline{\sigma}^h)d\Omega =0 \qquad (8)$$

Different forms of such projection have of course been implemented for a long time in many codes - one of the most commonly used being that of nodal averaging. Other possibilities are open using for instance weighted averaging or more accurately putting

$$\underline{P} = \underline{N} \qquad (9)$$

This projection is most effective [33] - and can be economically achieved by a simple iteration process [21,32]. With its use effectivity indices defined as

$$\theta = \frac{\text{predicted error}}{\text{actual error}} \qquad (10)$$

are in the range of 0.8-1.1 for most practical situations and estimates are reliable. Perhaps surprisingly, other forms of smoothing give comparable results [2].

We refer the reader to references [21,22,27] for a wide range of elements and problem in which such effectivity was studied. In this paper we shall only show this in some selected examples for which adaptive refinement is studied.

Other methods of obtaining σ^* can of course be included in the general methodology. For instance a stress smoothing using extrapolation from optimal sampling points in elements [34] has proved very effective in some problem using isoparametric elements [35].

The procedure is very effective and cheap for finite element up to the order p=2 but other possibilities of achieving the same end exist albeit involving some additional cost. Here in particular we should mention the technique based on residual projection used effectively by Oden, Demkowicz et al [19] and recent experiments involving the properties of superconvergence [31]. Such methods are probably necessary for p > 2 but as very many current elements are those in the lower range the estimator here outlined can be used with some confidence.

The use of energy norm estimators is particularly convenient of some overall percentage error

$$\eta = \frac{\|\underline{e}\|}{\|\underline{u}\|} \qquad (11)$$

is desired for comparative purpose and to guide the analysis. However from practical viewpoint more local estimates are often needed. Indeed these are available through the same analysis which computes energy norm errors element by element. Such local estimators can serve well as guides to local stress error magnitudes and again here show a surprisingly high effectivity [21].

The preceeding has, we believe, shown that in the error estimate field a state of art has been reached where its neglect can not be justified.

2.2 Prediction of necessary h-refinement for a specified accuracy.

The user will generally require that the overall accuracy η is at or below some specified value $\bar{\eta}$. It is unlikely that he will ever be satisfied with the results of a single analysis on a mesh designed 'using previously acquired knowledge' or experience. He will either overshoot the mark - and using overrefinement incur unnecessary analysis cost - or more frequently will find results which do not satisfy his initially specified criterion.

Two choice are then open to achieve desired results. He can either increase the order of trial functions p in the elements used initially (following so called p-refinement) or reduce the element size h used (h-refinement). Indeed he may try to vary both simultaneously with so called h-p process refinement. Much attention has been given to all of above but in the present practical sense he will generally have only a limited repertoire of possibilities open in a given code possessing a fixed element library. For this reasons in the present section we shall confine our attention to the h-refinement process alone and will show that after the initial analysis it is possible to predict an element size distribution needed for a specified accuracy. Indeed if such a size distribution is achieved by suitable remeshing (which we will show later) frequently a single reanalysis will suffice. An example of such an automatic process is shown in Fig. 4 where a single re-analysis allowed the goal of 5% accuracy to be reached from an initial analysis with an error about 17 per cent.

(FIGURE 4)

The mesh size prediction process is simple and aims towards the achievement of an optimal mesh for which the error in energy norm is equal in each element.

Thus if the actual error on a given mesh on which the original analysis was carried out is such that

$$\eta > \bar{\eta} \tag{12}$$

and we know $\|\underline{e}\|_m$, the energy norm of error associated with each element m, (m=1 to M), we shall first estimate for each of the existing elements the ratio

$$\xi_m = \frac{\|\underline{e}\|_m}{\|\underline{e}\|_{permissible}} \qquad (13)$$

where the permissible element error is given approximately (assuming an equal error distribution) as

$$\|\underline{e}\|_{permissible} \cong \bar{\eta}(\|\underline{u}^h\|^2 + \|\underline{e}\|^2)^{1/2} / \sqrt{M} \qquad (14)$$

observing here that only the square of the norm is additive.

Now we make use of the well know fact that [36,37]

$$\|\underline{e}\| \propto h^{\min(\lambda,p)} \qquad (15)$$

where λ is the 'strength' of singularities present, together with the fact that for an optimal mesh dependence on λ is eliminated. Though the mesh is not yet 'optimal' we can (with some optimism) predict a new element size required over the area of each present element m. Thus in each such subdomain we shall require

$$h^m_{new} = h^m_{existing} / \xi^{1/p} \qquad (16a)$$

However when considering the elements adjacent to the singularities it is desirable to predict the mesh size by

$$h^m_{new} = h^m_{existing} / \xi^{1/\lambda} \qquad (16b)$$

Practice has shown that this prediction is remarkably efficient and the code TAP-2D includes this simple feature which immediately gives the necessary guidance to the user for the refinement process.

Identical prediction is of course applicable to the problems of multiple loading. Now on the original mesh each of the loads load i (i=1,2,...,n) is applied, errors estimated and predicted mesh sizes $h^m_{new(i)}$ obtained by equation (16). The smallest predicted size at any location, i.e

$$h^m_{new} = \min(h^m_{new(1)}, h^m_{new(2)}, \ldots, h^m_{new(i)}, \ldots, h^m_{new(n)}) \qquad (17)$$

will be used to generate the final mesh and of course we would expect that this mesh would exceed slightly the accuracy required for each individual load.

2.3 Currently available procedures of automatic mesh refinement/Adaptivity.

To achieve the required mesh size distribution two main direction can be followed in principle; that of mesh enrichment in which the original mesh is retained and that of mesh regeneration in which either the whole of the mesh is redesigned or only the portion in which $\xi>1$.

(FIGURE 5)

Mesh enrichment, which generally precludes the possibility of de-refining i.e using a coarser mesh when $\xi<1$, is shown in Fig. 5. Here immediately we notice the incompatibilities which arise at element interfaces which generally involve the insertion of suitable, if complex, constrains and a special data structure [38,18]. For this reason the only practical enrichment procedure so far produced have involved successive halving of subdivision with many intermediate steps necessary to achieve final accura·y.

It appears that probably the partial or complete mesh regeneration processes are preferable as it is possible to achieve in a single operation the desired element size distribution predicted on the basis of equation 16. This is certainly the case where triangular subdivisions are used as those can be constructed using an advancing front technique to follow any specified size distributions [39]. Similar construction can of course be extended to three dimension and used to construct tetrahedral meshes at prescribed density [40]. Of course triangles and tetrahedrals can also be constructed using techniques developed by Shephard et al[41,42] and similar objective achieved. However with such methods it appears more difficult to follow the precise mesh density distributions.

Defining adaptivity as the process of adjustment to meet specified requirements following the examination of the present conditions we have shown in Fig.4 a typical procedure of refinement in which only a single step of adjustment was necessary. Later we shall show some further examples of such a fully automatic process derived by using a code MAD-2D (Mesh Adaptive Design) which follows the use of the previous post processor TAP-2D. Of course now some generality is lost as error estimates and mesh size prediction is applicable to almost all element shapes and types while the automatic mesh generation process is at the moment restricted to triangular (or tetrahedral) shapes. Indeed one of the unsolved problem is that of devising on automatic process of generating regular quadrilaterals, preferably on structured meshes, in which size distribution is prescribed.

Some generators based on various mapping procedure are being now investigated.One possibility is illustrated in Fig.6 in which triangular (or tetrahedral) meshes are generated first by procedures previously discussed and then subdivided into three quadrilaterals (or four hexahedral) shapes. However here the element shapes so derived are not to everybody's liking.

(FIGURE 6)

2.4 Some further examples of automatic adaptivity. Multiple loads, Thermal problems.

In the proceeding we have illustrated how a fully automatic procedure for achieving error estimators, mesh size prediction and results of a specified accuracy can be obtained today. Fig.4 as well as numerous examples in references [21-23,25-28,30] show possible application and their effectivity in various problems at linear and nonlinear analysis. In this section we show two additional examples illustrating the application of the procedure to slightly different situation.

(FIGURE 7)

The first shows a typical adaptive process in plane stress elastic analysis for multiple loads Fig.7. Here the application of the error analysis programs is made independently for each load using the original mesh and the desired mesh sizes to obtain 5% (with the use of quadratic elements) and 15% accuracy (with the use of linear elements) in each load case are computed. Further a single final mesh on which each load is again applied separately is generated leading of course to higher accuracy then that original specified but reducing the computational cost of generating separate meshes and solving each load case individually.

For comparison both types of solution are included in Fig. 7 with linear and quadratic elements.

(FIGURE 8)

The second example considers a problem of heat conduction and thermal stresses in which first an adaptively refined mesh for temperature solution is obtained following by a second thermal stress solution. Quadratic elements are used in the analysis.The prescribed accuracy of 5% has been achieved for both temperature solution and thermal stress solution on refined mesh. Fig. 8.

3. Error estimation and adaptivity -- What will be available tomorrow.

So far we have discussed methodologies which can easily be adopted to fairly standard existing codes widely used in practice without necessity for major restructuring. With such approaches accuracies of 3-5% are easy to reach without excessive refinement. For higher precision p- or h-p refinement processes are necessary but here major changes of codes are needed and doubtless in the future this option will be incorporated in many programs. The use of local hierarchical p refinement firstly introduced in 1971 [43] was successfully used by Peano, Szabo et al [44-47]. Recently Oden et al [18-20] have shown how combination of such p-refinement with h enrichment can result in a minimum number of parameters required to reach a given accuracy (this is also reported elsewhere in this issue).

(FIGURE 9)

The combination of h and p process can however be accomplished with better computational economy by use of a simple h-refinement leading to say a 5% accuracy followed by a uniform increase of p over the whole domain [29]. With the first refinement carried out using p=2 it is found that an increase of p to 4 will generally lead to energy norm accuracy better than 1%. A typical example is illustrated in Fig. 9a. An improvement of this procedure uses a higher order p to predict the mesh size for the h-refinement. Now a better mesh and more rapid convergence will be obtained.(Fig. 9b)

In uniform p refinement error estimation is quite accurate and easy to accomplish providing results for <u>three</u> values of p are available. With the error given as

$$\|\underline{e}\| = (\|\underline{u}\|^2 - \|\bar{\underline{u}}\|^2)^{1/2} = CN^{\alpha} \qquad (18)$$

where N is the number of degrees of freedom, and C,α are constants, three solutions with different valuas of p yield $\|\underline{u}\|$, C and α and therefore $\|\underline{e}\|$.

Of more importance is the extension of error estimation and adaptivity to non-linear and transient problems. Much current research is in progress in both areas. In Fig. 10 a highly non-linear and transient problem of metal forming is illustrated [28]. Here error estimation and mesh refinement follow precisely the line used in section 2 but to allow for less frequent refinement than that at each step two limits for η are used. An upper value $\bar{\bar{\eta}}$ which must not be exceeded and $\bar{\eta}$ set lower for which we aim when the upper value is exceeded.

(FIGURE 10)

A problem of considerable interest in current studies of non-linear behavior is that of localization which occurs under certain conditions of plastic behavior or fracture. In this area again local refinement following error estimating procedures is of great importance and the subject is currently under active research. Fig. 11 [30] shows how such localization can be clearly indicated by an h-refinement process.

(FIGURE 11)

In the same Figure we show how the use of elongated elements can produce same accuracy from fewer degrees of freedom in localization which is essentially one dimensional.

(FIGURE 12)

This concept is of great importance in hyperbolic prol lem of fluid mechanics of compressible flow where indeed it was first introduced [39-40]. Here shock localization and capturing can be effectively performed using this device Fig. 12, however the error measure used in such problem are different. It is probably that the greatest strides in tomorrow's developments will indeed come from this important area in which three dimensional mesh generation and refinement were indeed first introduced Fig.13.

(FIGURE 13)

REFERENCES

1. I.Babuska and T.Scapolla, 'Benchmark computation and performance evaluation for a rhombic plate bending problem', *Int. J. Numer. Meth.Eng.* **28**, (1989), 155-179.

2. O.C.Zienkiewicz and J.Z.Zhu, 'Error estimates and adaptive refinement for plate bending problems', *Int. J. Numer. Meth. Eng.* **28**, (1989), 2839-2853.

3. R.L.Taylor, O.C.Zienkiewicz, J.C.Simo and A.H.C.Chan, 'The patch test - condition for assessing FEM convergence', *Int. J. Numer. Meth. Eng.* **22**, (1986), 32-62.

4. O.C.Zienkiewicz, S.Qu, R.L.Taylor and S.Nakazawa, 'The patch test for mixed formulations', *Int. J. Numer. Meth. Eng.*, **23**, (1986), 1873-1883.

5. O.C.Zienkiewicz and D.Lefebvre, 'Three field mixed approximation and the plate bending problems', *Comm. Appl. Numer. Meth.*, **3**, (1987), 301-309.

6. I.Babuska, 'The finite element method with Lagrange multipliers', *Numer. Math.*, **20**, (1973), 179-192.

7. F.Brezzi, 'On the existence, uniquence and approximation of saddle point problems arising from Lagrange multipliers', RAIRO, 8(r-2), (1974), 127-151.

8. I.Babuska and W.C.Rheinboldt, 'A posteriori error estimates for the finite element method', *Int. J. Numer. Meth. Eng.*, 12, (1978), 1597-1615.

9. I.Babuska and W.C.Rheinboldt, 'Error estimates for adaptive finite computations' *SIAM. J. Numer. Anal*, 15, (1978), 736-754.

10. D.W.Kelly, J.P.R.Gago, O.C.Zienkiewicz and I.Babuska, 'A posteriori error analysis and adaptive processes in the finite element method', Part I, *Int. J. Numer. Meth. Eng.*, 19, (1983), 1593-1619.

11. J.P.R.Gago, D.W.Kelly, O.C.Zienkiewicz and I.Babuska, 'A posteriori error analysis and adaptive processes in the finite element method', Part II, *Int. J. Numer. Meth. Eng.*, 19, (1983), 1621-1656.

12. O.C.Zienkiewicz, J.P.R.Gago and D.W.Kelly, 'The hierarchic concept in finite ement analysis', *Computers and Structures*, 16, No.14, (1983), 53- 5.

13. I.Babuska and A.Miller, 'The post-processing approach in the finite element method - Part 3: A-posteriori error estimates and adaptive mesh selection', *Int. J. Num. Meth. Eng.*, 20, (1984), 2311-2324.

14. R.E.Bank and A.Weiser, 'Some a posteriori error estimates for

elliptic partial differential equations', *Math. Comp.* **44**, (1985), 283-301.

15. R.E.Bank, 'Analysis of a local a posteriori error estimate for elliptic equations', Chapter 7, 119-128, Accuracy Estimates and Adaptive Refinements in Finite Element Computations, John Wiley & Sons, 1986.

16. J.T.Oden, L.Demkowicz, T.Strouboulis and P.Devloo, 'Adaptive methods for problems in solid and fluid mechanics', Chapter 14, 249-280, Accuracy Estimates and Adaptive Refinements in Finite Element Computations, John Wiley & Sons, 1986.

17. I.Babuska, O.C.Zienkiewicz, J.Gago and E.R.De.Oliveira (eds.), Accuracy Estimates and Adaptive Refinements in Finite Element Computations, Wiley, 1986.

18. L.Demkowicz, J.T.Oden, W.Rachowicz and O.Hardy, 'Toward a universal h-p adaptive finite element strategy. Part 1. Constrained approximation and data structure', *Computer Meth. Appl. Mech. Eng.*, 77, (1989), 79-112.

19. J.T.Oden, L.Demkowicz, W.Rachowicz and T.A.Westermann, 'Toward a universal h-p adaptive finite element strategy. Part 2. A posteriori error estimation', *Computer Meth. Appl. Mech. Eng.*, 77, (1989), 113-180.

20. W.Rachowicz, J.T.Oden and L.Demkowicz, 'Toward a universal h-p

adaptive finite element strategy. Part 3. Design of h-p meshes', *Computer Meth. Appl. Mech. Eng.,* 77, (1989), 181-212.

21. O.C.Zienkiewicz and J.Z.Zhu, 'A simple error estimator and adaptive procedure for practical engineering analysis', *Int. J. Num. Meth. Eng.*, 24, 337-357, 1987.

22. O.C.Zienkiewicz, J.Z.Zhu, Y.C.Liu, K.Morgan and J.Peraire, 'Error estimates and adaptive from elasticity to high speed compressible flow', *MAFELAP 87, ed. J.R. Whiteman, Academic Press,* (1988), 483-512.

23. J.Z.Zhu and O.C.Zienkiewicz, 'Adaptive techniques in the finite element method', *Comm. Appl. Meth. Eng.* 4, (1988), 197-204.

24. E.Rank and O.C.Zienkiewicz, 'A simple error estimator in the finite element method', *Comm. Appl. Num. Math.,* 3, (1987), 243-249.

25. M.Ainsworth, J.Z.Zhu, A.W.Craig and O.C.Zienkiewicz, 'Analysis of the Zienkiewicz-Zhu a-posteriori error estimator in the finite element method', *Int. J. Numer. Meth. Eng.*, 28, (1989), 2161-2174.

26. O.C.Zienkiewicz, Y.C.Liu, and G.C.Huang, 'Error estimation and adaptivity in flow formulation for forming problem', *Int. J. Numer. Meth. Eng,* 25, (1988), 23-42.

27. O.C.Zienkiewicz, Y.C.Liu and G.C.Huang, 'Error estimates and

convergence rates for various incompressible elements', *Int. J. Numer. Meth. Eng.*, 28, (1989), 2191-2202.

28. O.C.Zienkiewicz, G.C.Huang and Y.C.Liu, 'Adaptive FEM computation of forming processes - Application to porous and nonporous materials', *Int. J. Numer. Meth. Eng.* to be published, 1990.

29 O.C.Zienkiewicz, J.Z.Zhu and N.G.Gong, 'Effective and practical h-p version adaptive analysis provedures for the finite element method', *Int. J. Numer. Meth. Eng.*, 28, (1989), 879-891.

30. O.C.Zienkiewicz and G.C.Huang, 'A note on localization phenomena abd adaptive finite element analysis in forming processes', *Comm. Appl. Num. Meth.*, to appear, 1990.

31. J.Z.Zhu and O.C.Zienkiewicz, 'Superconvergence recovery technique and a-posteriori error estimators', *Int. J. Numer. Meth. Eng.* to be published, 1990.

32. O.C.Zienkiewicz and R.L.Taylor, The Finite Element Method, Fourth edition, McGraw-Hill, 1989.

33. J.T.Oden and H.J.Brauchli, 'On the calculation of consistent stress distributions in finite element applications', *Int. J. Num. Meth. Eng.*, 3, (1971), 317-325.

34. E.Hinton and J.S.Campbell, 'Local and global smoothing of

discontinuous finite element functions using a least squares method', *Int. J. Numer. Meth. Eng.* 8, (1974), 461-480.

35. H.Pircher, Private Communication, TDV, Gratz, Austia, 1987.

36. P.G.Ciarlet, The Finite Element Method for Elliptic Problems, Amsterdan: North-Holland, 1978.

37. J.T.Oden and G.F.Carey, Finite Elements: Mathmatical Aspects, Vol. IV. Englewood Cliffs, N.J.: Prentice-Hall, 1983.

38. G.F.Carey, M.Sharma and K.C.Wang, 'A class of data structures for 2-D and 3-D adaptive mesh refinement', *Int. J. Num. Meth. Eng.*, 26, (1988), 2607-2622.

39. J.Peraire, M.Vahdati, K.Morgan and O.C.Zienkiewicz, 'Adaptive remeshing for compressible flow computations', *J. Computational Physics, vol.* 72, *no.* 2, (1987), 449-466.

40. J.Peraire, J.Peiro, L.Formaggia, K.Morgan and O.C.Zienkiewicz, 'Finite element Euler computations in three dimensions', *Int. J. Num. Meth. Eng.*, 26, (1988), 2135-2159.

41. M.S.Shephard, P.L.Baehmann and K.R.Grice, 'The versatility of automatic mesh generators based on tree structures and advanced geometric constructs', *Comm. Appl. Numer. Meth.* 4, (1988), 379-392.

42. P.L.Baehmann, S.L.Wittchen, M.S.Shephard, K.R.Grice and

M.A.Yerry, 'Robust, geometrically based, automatic two-dimensional mesh generation', *Int. J. Num. Meth. Eng.*, 24, (1987), 1043-1078.

43. O.C.Zienkiewicz, B.M.Irons, F.E.Scott and J.S.Campbell, 'High speed computing of elastic structures', *Proc. Symp. Int. Union of Theoretical and Applied Mechanics*, Liege, 1970.

44. A.G.Peano, 'Hierarchies of conforming finite elements for plane elasticity and plate bending', *Comput. and Maths. with Appls.* 2, (1976), 211-224.

45. A.G.Peano, A.Pasini, R.Riccioni and Sardella, 'Adaptive approximation in finite element structural analysis', *Computers and Structures*, 10, 332-342, 1979.

46. B.A.Szabo, 'Some recent developments in finite element analysis', *Comput. and Maths. with Appls.* 5, 99-115, 1979.

47. A [illegible]bo. 'Mesh design for the p-version of the finite element method', *Computer Meth. Appl. Mech. Eng.*, 55, 181-197, 1986.

Figure 1. A two dimensional half space modelling a foundation. How big should depth d of the analysis region be to determine realistically displacement of structure under load W ?

Figure 2. Thick and thin plate problems.
What is a simple support? How to deal with local 3D support displacement? What are plate spans? Is it a plate or a membrane?

Figure 3. Rounded and sharp notches in a torsion bar.
How does the stress vary as notch size decreases to zero? What size of surface imperfection can be ignored?

Figure 4. Plane strain problem of a dam under water load. Automatic adaptive mesh generation to achieve 5% accuracy.

Figure 5. Difficulties of mesh enrichment to specified element size (without constrained data structure).

Figure 6. Quadrilaterals and hexahedrals (bricks) generated by subdivision of triangles or tetrahedra.

Figure 7a. Adaptive analysis for two loads separate and combined meshes aiming for 5% accuracy with quadratic element.

Figure 7b. Adaptive analysis for two loads separate and combined meshes aiming for 15% accuracy with linear element.

Figure 8. Automatic adaptive analysis of heat conduction and thermal stress (plane strain) problem to achieve 5% accuracy. Quadratic triangular element.

Figure 9a. Procedure 1 in the h-p version analysis of short cantilever beam with rigidly fixed side. Quadratic triangular element is first used in h version to achieve 5% accuracy. p is subsequently increased up to 4 to achieve 1% accuracy.
(a). Original mesh. (b). h-refined mesh. (c). h and p refined mesh. (d). Convergence of results with uniform refinement and h-p adaptive procedure.

Figure 9b. Procedure 2 in the h-p version analysis of short cantilever beam with rigidly fixed side. p refinement used in original mesh (a) achieving 11.8% accuracy at p=4. This allows error in each element to be determined for p=3 (b). Final convergence of results shown in (c).

Figure 10. Two stages of extrusion of metal through a tapered die.
(a). Material grid.
(b). Mesh before remeshing.
(c). Mesh after remeshing.

Figure 11. Compression of a uniform, ideally plastic specimen with a central defect modelled as a circular opening.

Figure 12. Directional mesh refinement. Flow past a circular cylinder -- Mach 3. 3rd. refinement mesh 709 nodes (1348 elements).

Figure 13. An adaptive three dimensional analysis of compressible flow around an aircraft. Mesh of tetrahedral elements and contour plots of pressure distribution.

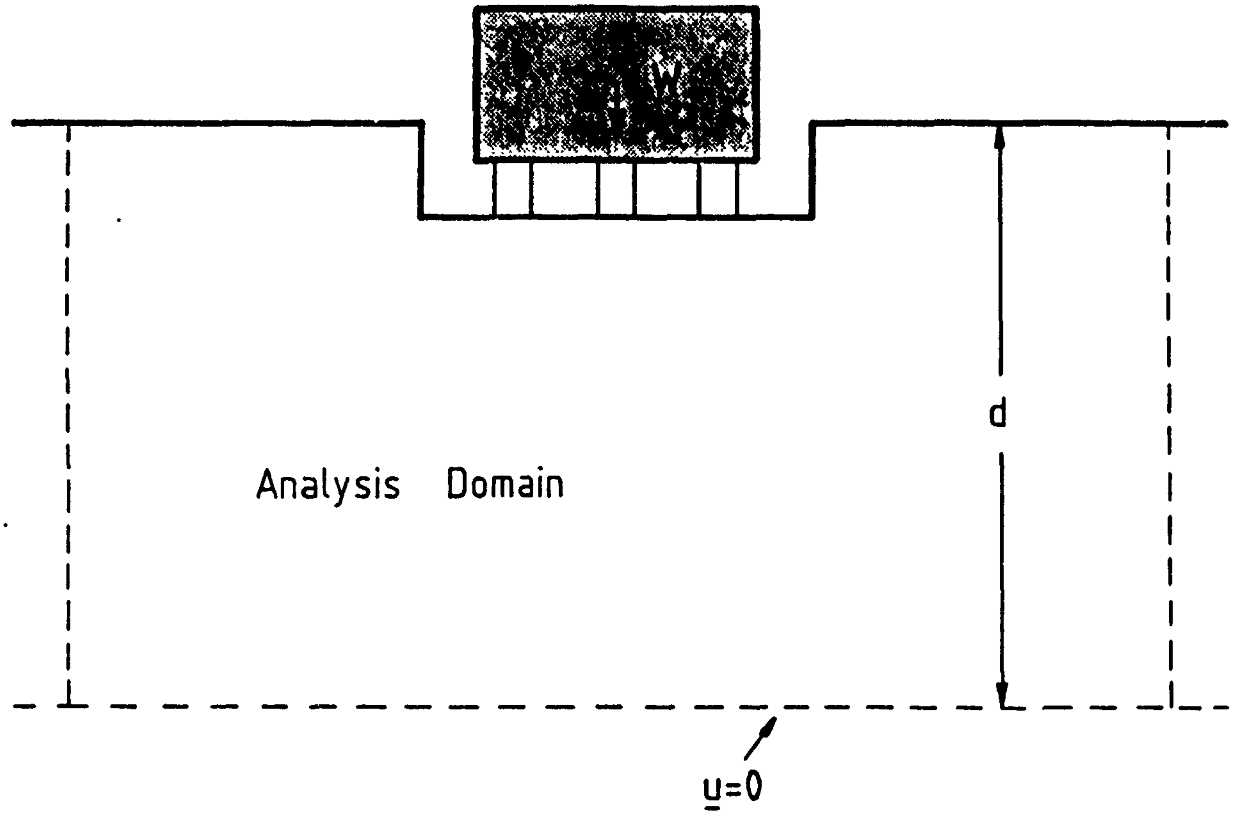


FIGURE 1. A TWO DIMENSIONAL HALF SPACE MODELLING A FOUNDATION. HOW BIG SHOULD DEPTH d OF THE ANALYSIS REGION BE TO DETERMINE REALISTICALLY DISPLACEMENT OF STRUCTURE UNDER LOAD W ?

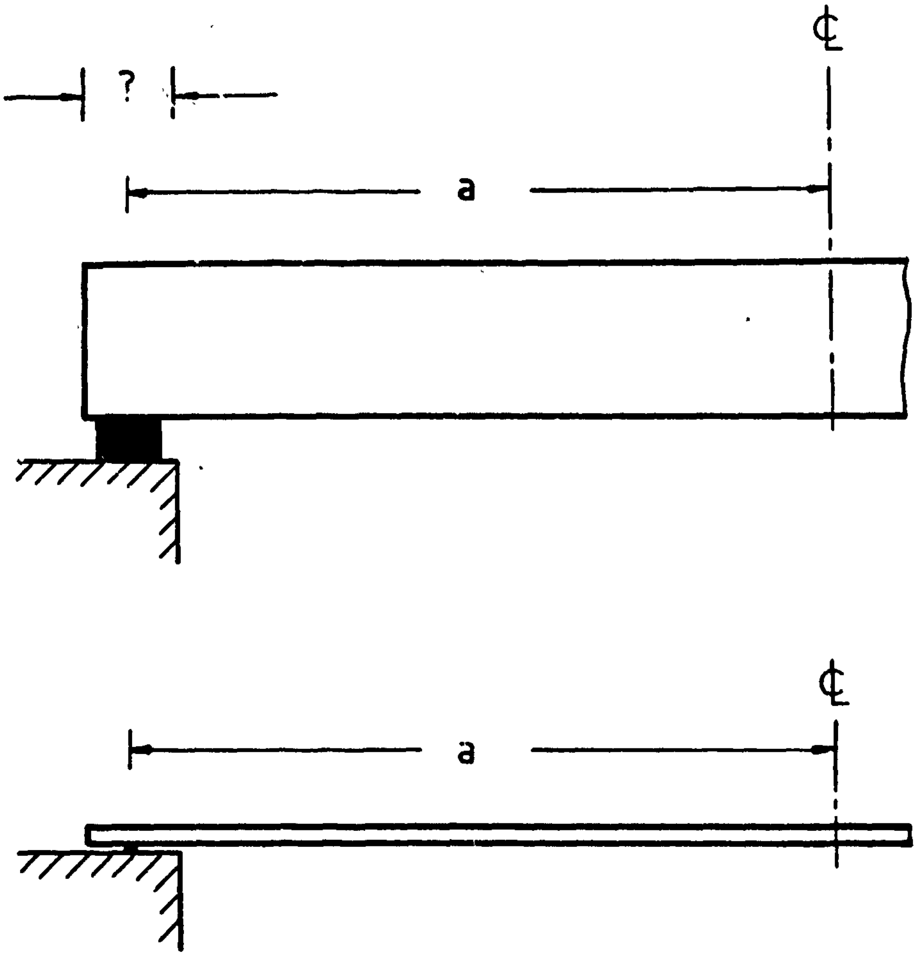


FIGURE 2 THICK AND THIN PLATE PROBLEMS.

What is a simple support ?
How to deal with local 3D support displacements
What are plate spans ?
Is it a plate or a membrane ?

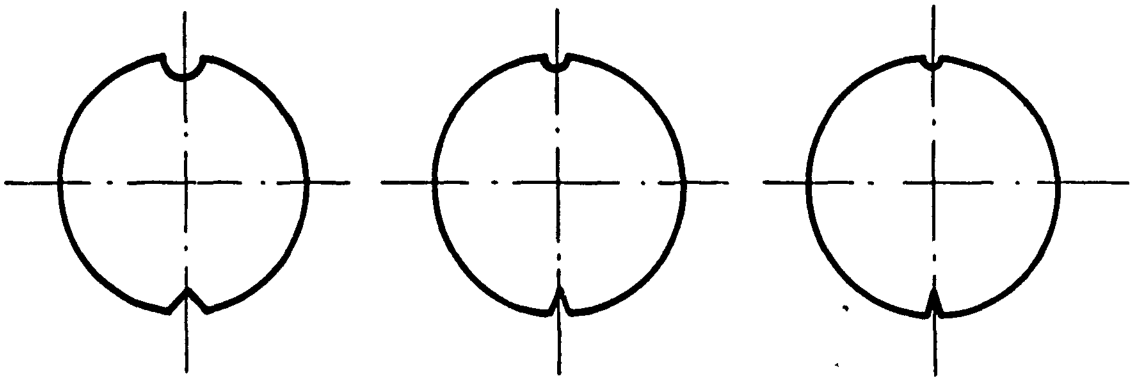
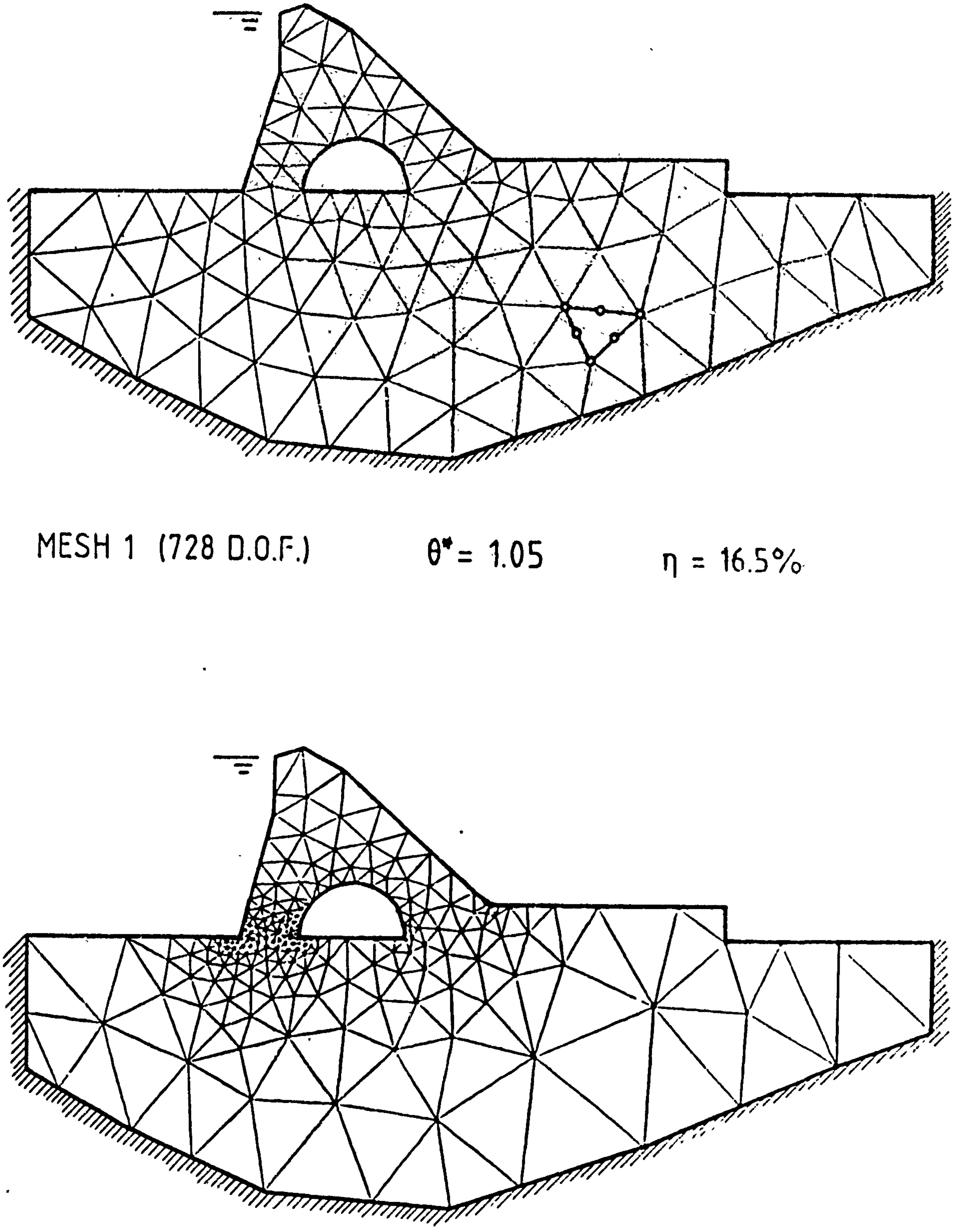


FIGURE 3 ROUNDED AND SHARP NOTCHES IN A TORSION BAR

How does the stress vary as notch size decreases to zero ?
What size of surface imperfection can be ignored ?



MESH 1 (728 D.O.F.) $\theta^* = 1.05$ $\eta = 16.5\%$

MESH 2 (1764 D.O.F.) $\theta^* = 1.07$ $\eta = 4.88\%$

FIGURE 4 PLANE STRAIN PROBLEM OF A DAM UNDER WATER LOAD. AUTOMATIC ADAPTIVE MESH GENERATION TO ACHIEVE 5% ACCURACY

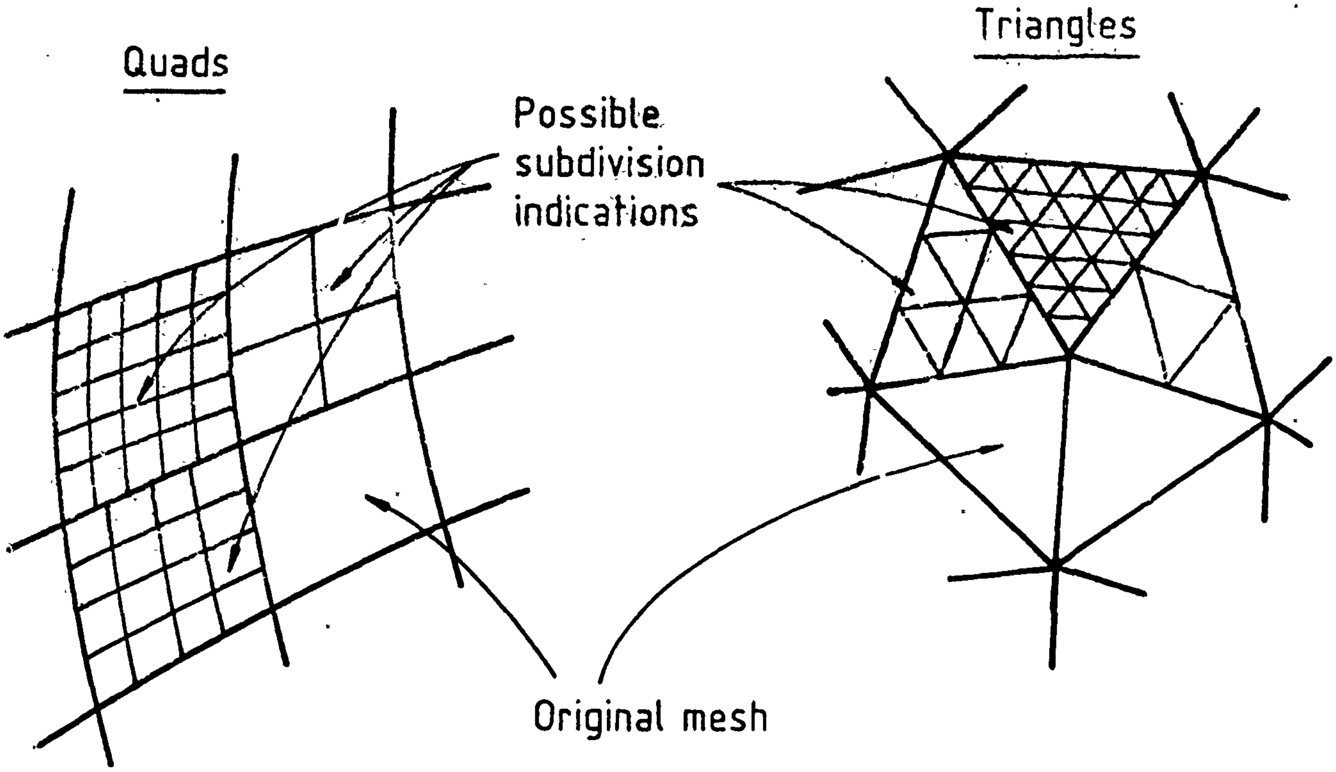


FIGURE 5 DIFFICULTIES OF MESH ENRICHMENT TO SPECIFIED ELEMENT SIZE (WITHOUT CONSTRAINED DATA STRUCTURE)

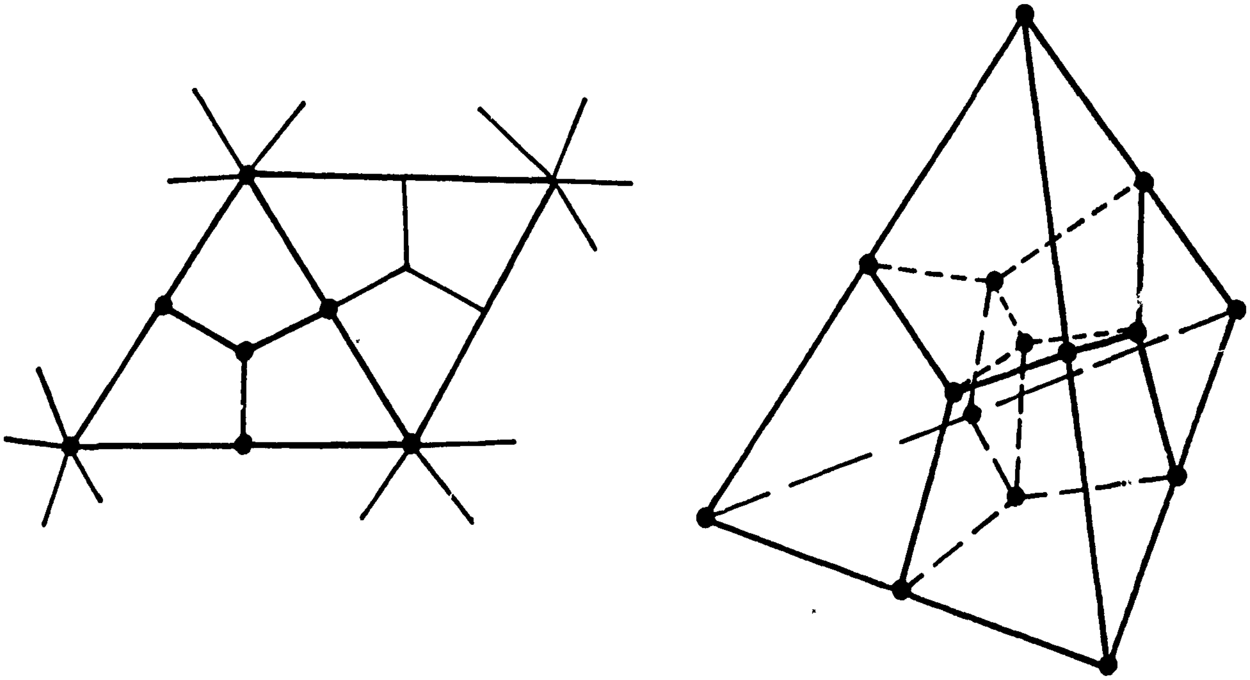
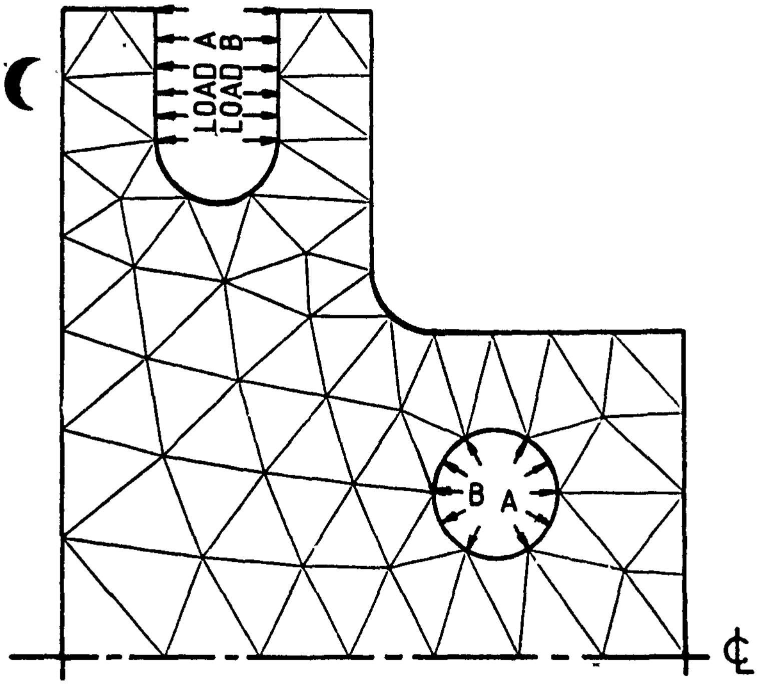
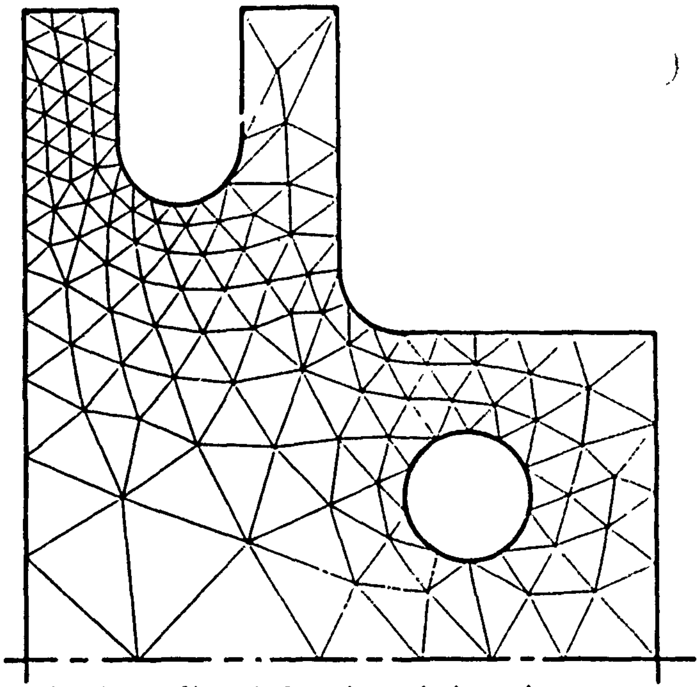


FIGURE 6 QUADRILATERALS AND HEXAHEDRALS (BRICKS) GENERATED BY SUBDIVISION OF TRIANGLES OR TETRAHEDRA

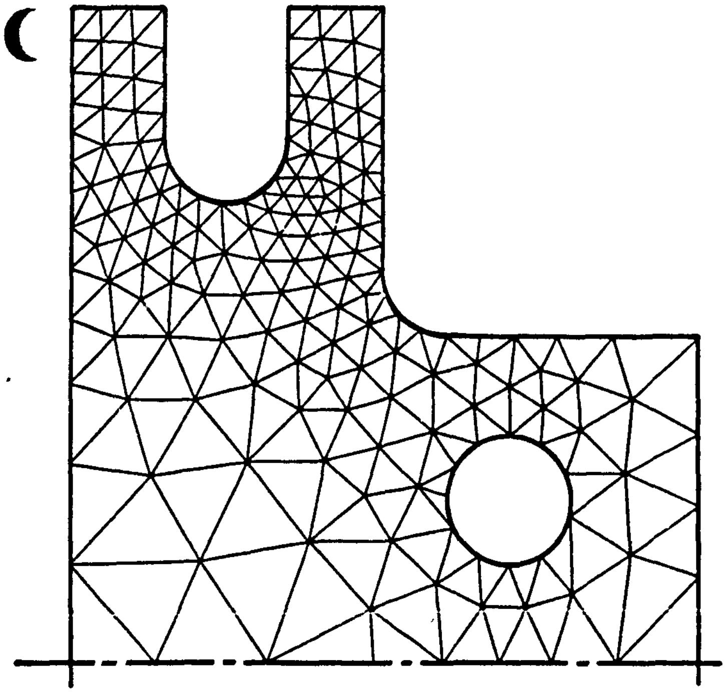
SEPARATE AND COMBINED MESHES
AIMING FOR 5% WITH QUADRATIC ELEMENT



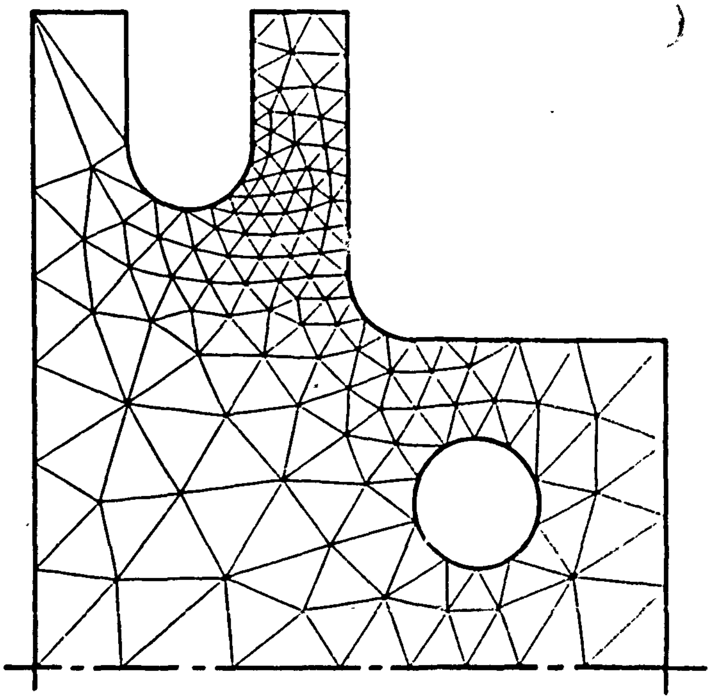
Original mesh (390 D.O.F.)
Load A η=11.58%
Load B η=13.82%



Mesh refined for Load A only
η=4.00% (1016 D.O.F.)

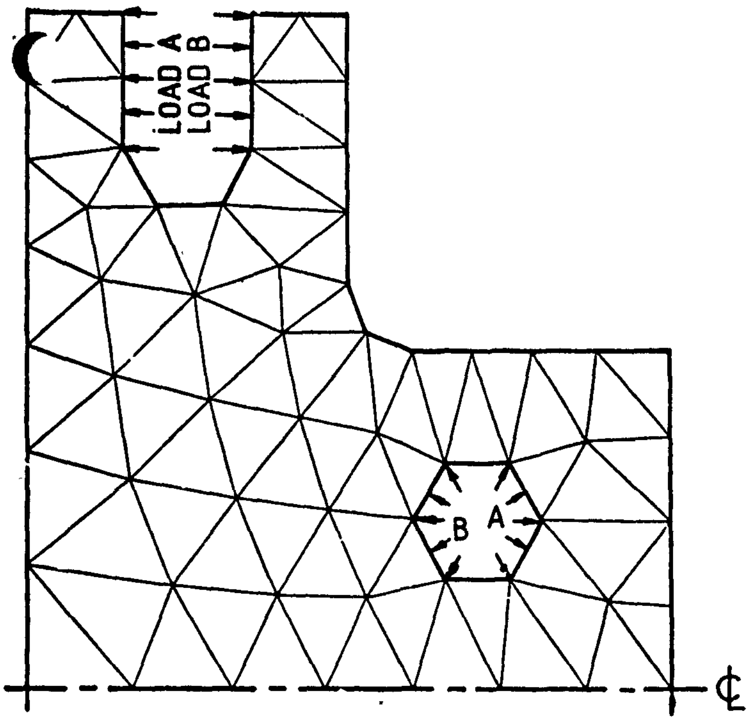


Mesh refined for Load A and B
Load A η=3.56% (1322 D.O.F.)
Load B η=3.60%

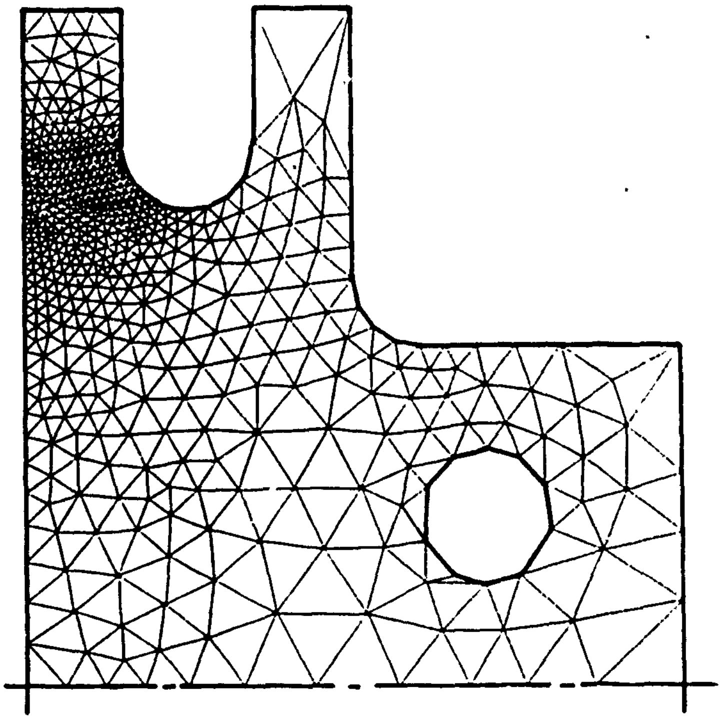


Mesh refined for Load B only
η=4.02% (1032 D.O.F.)

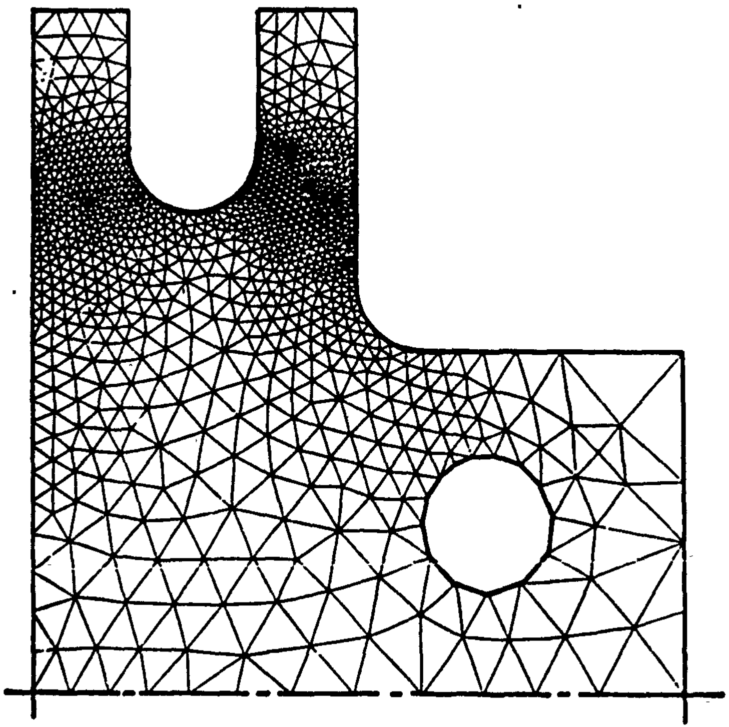
SEPARATE AND COMBINED MESHES
AIMING FOR 15% WITH LINEAR ELEMENT



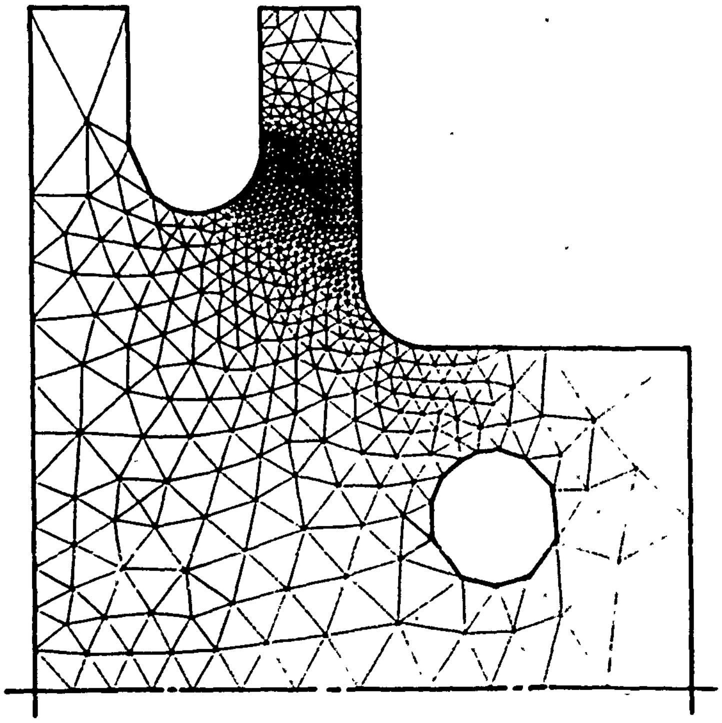
Original mesh (115 D.O.F.)
Load A η=50.21%
Load B η=51.78%



Mesh refined for Load A only
η=16.24% (1063 D.O.F.)

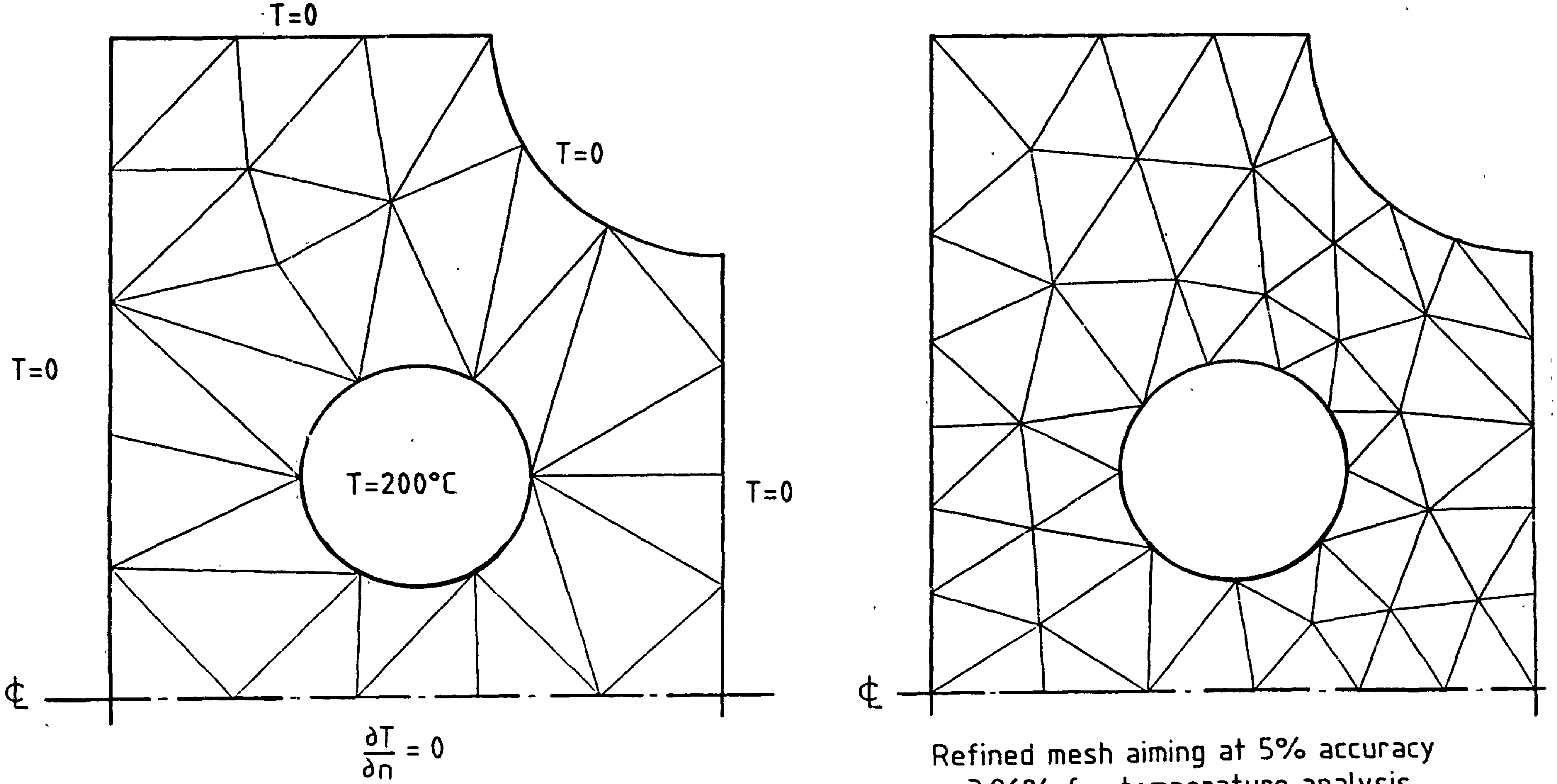


Mesh refined for Load A and B
Load A η=14.71% (1998 D.O.F.)
Load B η=14.16%



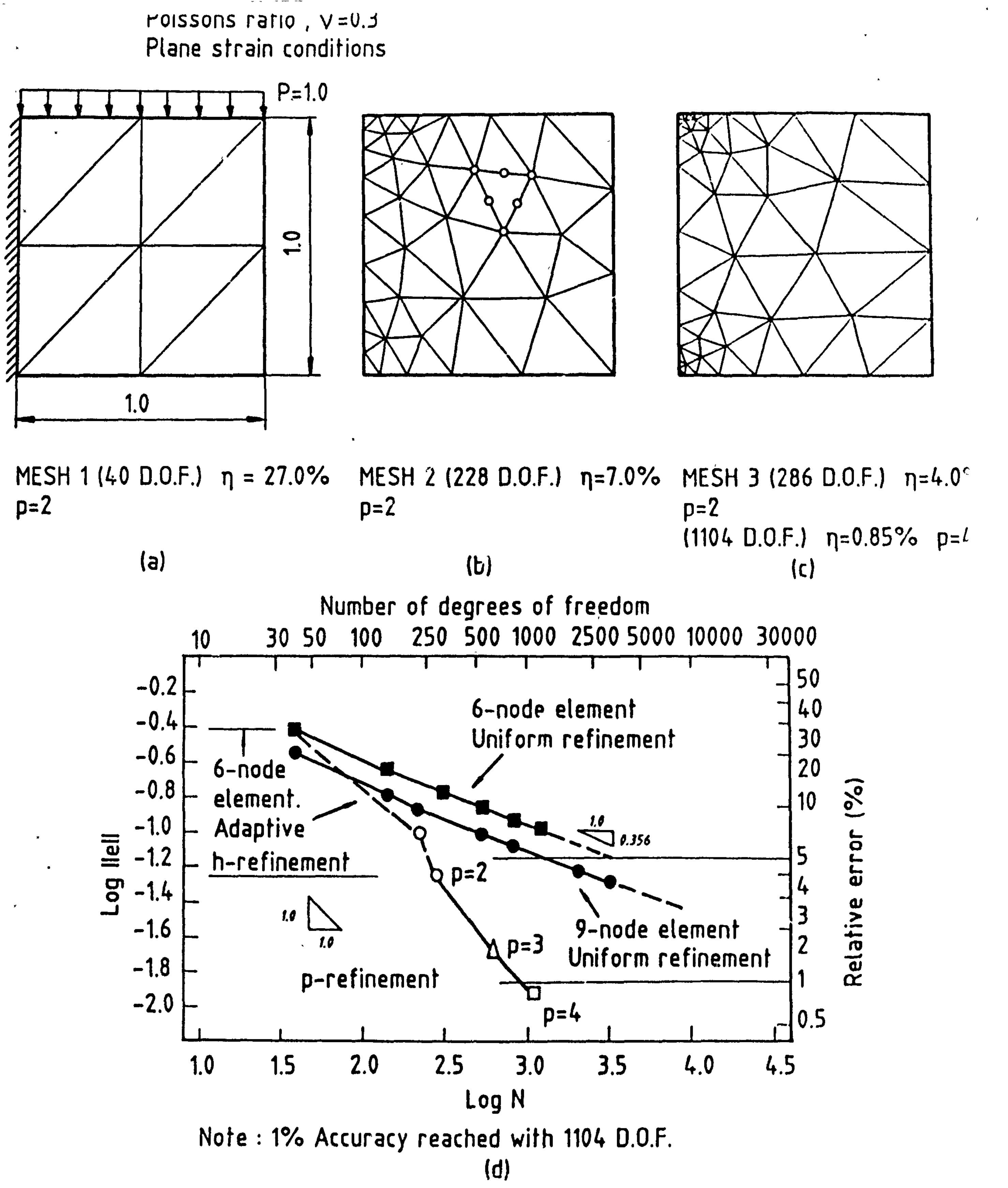
Mesh refined for Load B only
η=15.27% (1280 D.O.F.)

FIGURE 8 AUTOMATIC ADAPTIVE ANALYSIS OF HEAT CONDUCTION AND THERMAL STRESS (PLANE STRAIN) PROBLEM TO ACHIEVE 5% ACCURACY . QUADRATIC TRIANGULAR ELEMENT .



First mesh for temperature analysis
$\eta=9.98\%$

Refined mesh aiming at 5% accuracy
$\eta=3.96\%$ for temperature analysis
$\eta=2.71\%$ for thermal stress analysis



MESH 1 (40 D.O.F.) η = 27.0%
p=2

(a)

MESH 2 (228 D.O.F.) η=7.0%
p=2

(b)

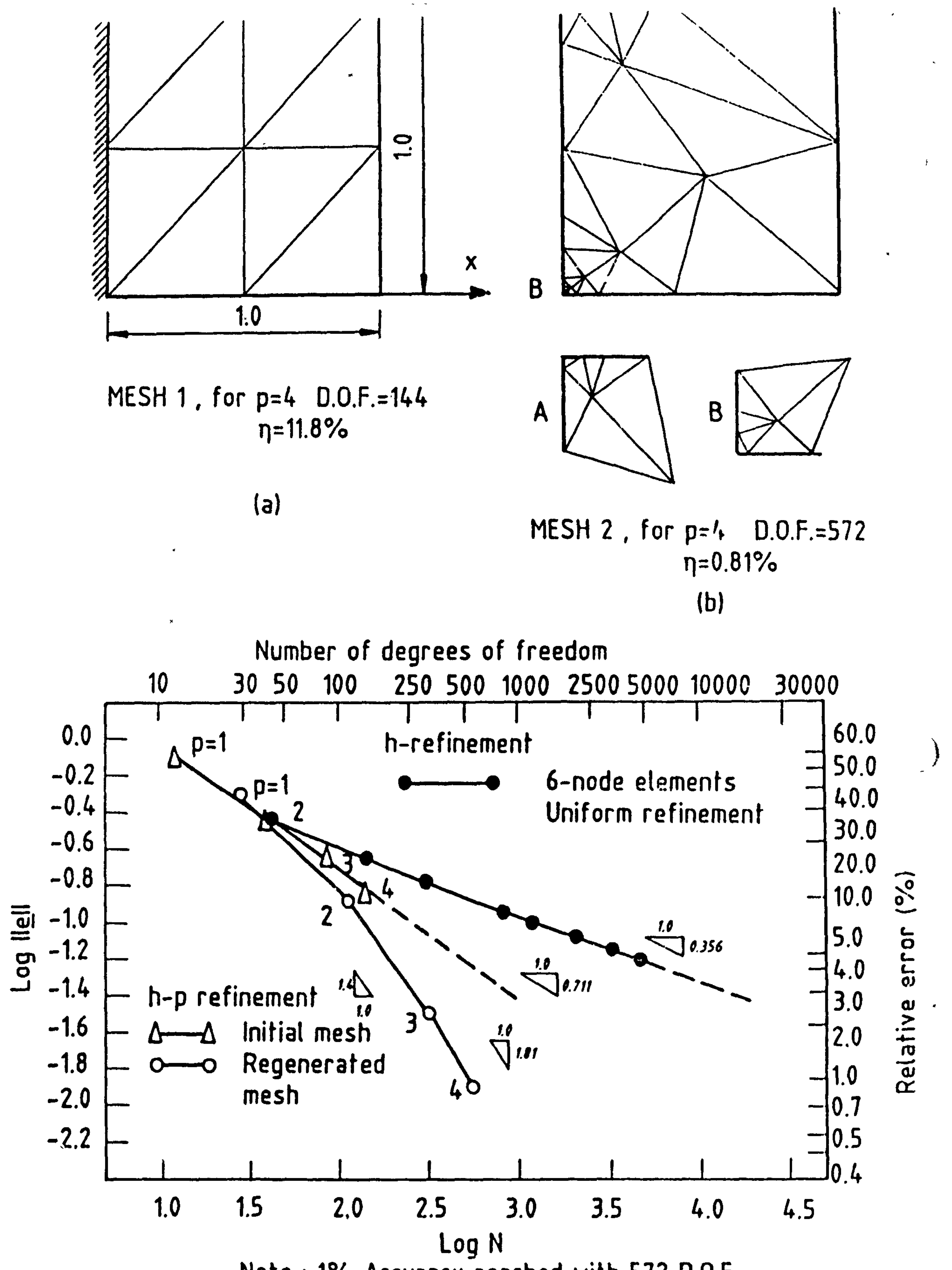
MESH 3 (286 D.O.F.) η=4.0
p=2
(1104 D.O.F.) η=0.85% p=

(c)

Note : 1% Accuracy reached with 1104 D.O.F.

(d)

FIGURE 9a PROCEDURE I IN THE h-p VERSION ANALYSIS OF SHORT CANTILEVER BEAM WITH RIGIDLY FIXED SIDE . QUADRATIC TRIANGULAR ELEMENT IS FIRST USED IN h VERSION TO ACHIEVE 5% ACCURACY . p IS SUBSEQUENT INCREASED UP TO 4 TO ACHIEVE 1% ACCURACY.
(a) Original mesh , (b) h Refined mesh , (c) h Refined mesh
(d) Convergence of results with uniform refinement and h-p adaptive procedure



MESH 1 , for p=4 D.O.F.=144
η=11.8%

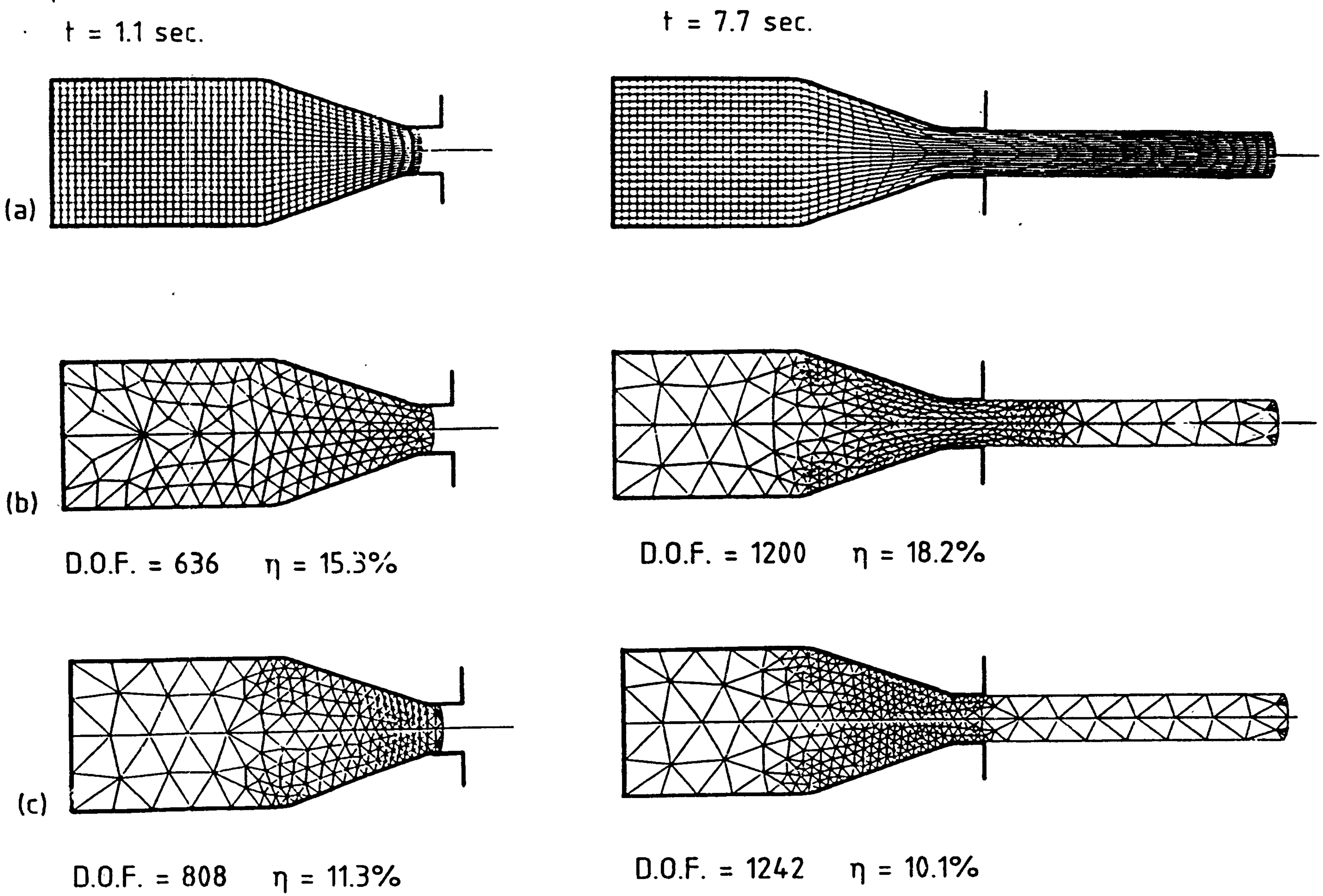
(a)

MESH 2 , for p=4 D.O.F.=572
η=0.81%

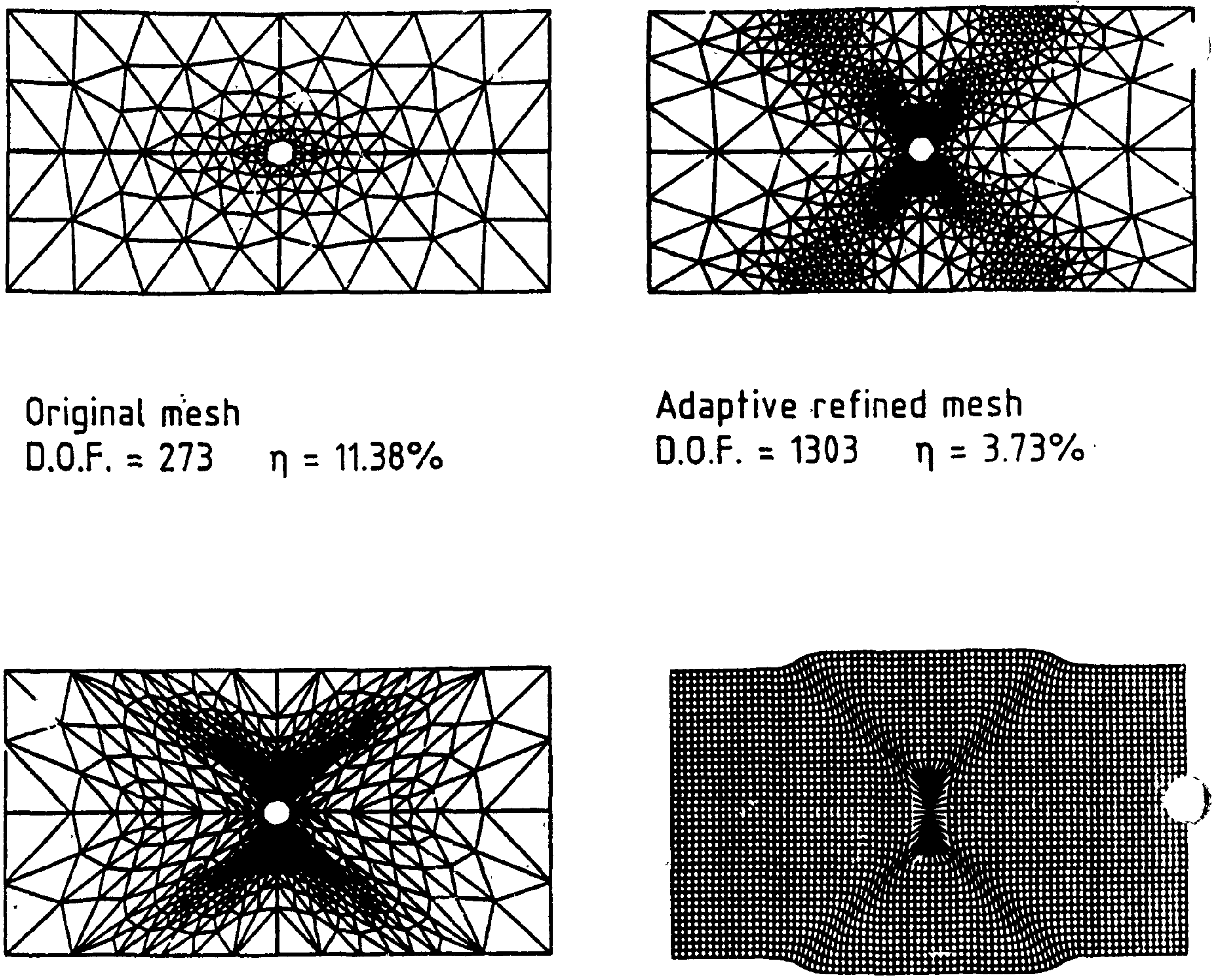
(b)

Note : 1% Accuracy reached with 572 D.O.F.

FIGURE 9b PROCEDURE II IN THE h-p VERSION ANALYSIS OF SHORT CANTILEV BEAM WITH RIGIDLY FIXED SIDE . p REFINEMENT USED IN ORIGINAL MESH (a) ACHIEVING 11.8% ACCURACY AT p=4. THIS ALLOWS ERROR IN EACH ELEMENT TO BE DETERMINED FOR p=3 (b) . FINAL CONVERGENCE OF RESULTS SHOWN IN (c)



4 FIGURE 10 TWO STAGES OF EXTRUSION OF METAL THROUGH A TAPERED DIE .
(a) Material grid (b) Mesh before remeshing (c) Mesh after remeshing



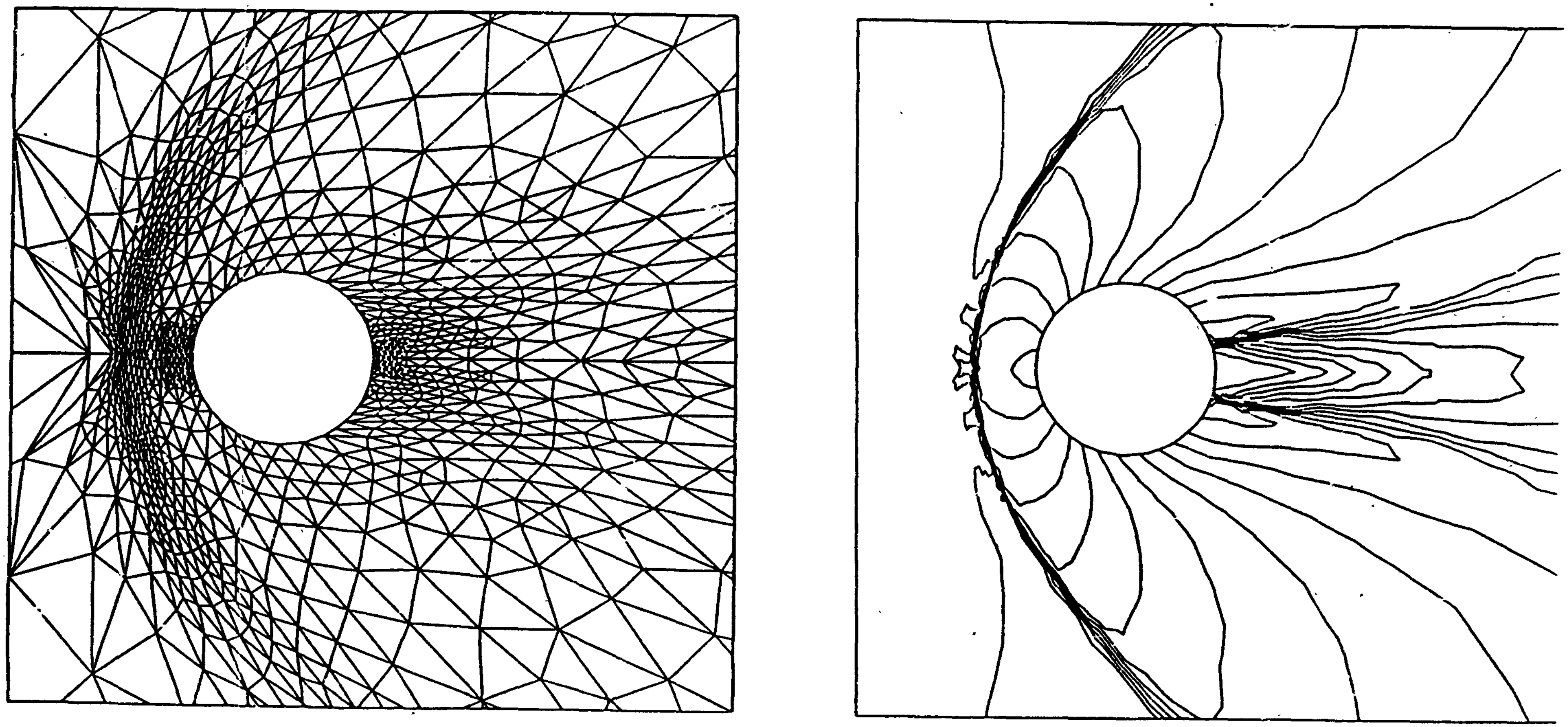
Original mesh
D.O.F. = 273 η = 11.38%

Adaptive refined mesh
D.O.F. = 1303 η = 3.73%

Adaptive refined elongated mesh
D.O.F. = 1039 η = 2.76%

Final deformed material grid

FIGURE 11 COMPRESSION OF A UNIFORM , IDEALLY PLASTIC SPECIMEN WITH A CENTRAL DEFECT MODELLED AS A CIRCULAR OPENING .



(a) Local mesh

(b) Pressure coefficients

FIGURE 12 DIRECTIONAL MESH REFINEMENT
FLOW PAST A CIRCULAR CYLINDER - MACH 3
3rd. REFINEMENT MESH 709 NODES (1348 ELEMENTS)

FACTORS AFFECTING RELIABILITY OF COMPUTER SOLUTIONS WITH HIERARCHICAL SINGLE SURFACE CONSTITUTIVE MODELS

C. S. DESAI, G. W. WATHUGALA, K. G. SHARMA AND L. WOO
Department of Civil Engineering and Engineering Mechanics
University of Arizona
Tucson, AZ 85721, U.S.A.

ABSTRACT

Influence and sensitivity of various material constants in the advanced hierarchical single surface (HISS) plasticity based models on finite element computer solutions are first discussed. Such factors as drift correction and time integration schemes in conjunction with various versions in the HISS approach are considered with respect to their influence on reliability and robustness of computer solutions. Practical examples involving dynamics of piles in porous anisotropically hardening soils, and static and dynamic response of concrete experiencing damage and softening are presented.

1. INTRODUCTION

Reliability and robustness of computational methods are dependent upon various factors related to the mathematical and numerical characteristics of the methods, properties of computers used, and various physical characteristics such as nonlinear material response and geometry.

1.1 Scope

Nonlinear material response is one of the vital factors that can influence significantly the reliability of computer solutions. Another important factor is appropriateness of the time integration schemes for nonlinear dynamics and field problems.

The main objective of this paper is to discuss a number of important attributes of material response, and a brief description of time integration schemes in the context of the nonlinear response.

A number of constitutive models, linear elastic, nonlinear elastic, classical elasto-plastic with unique description of yield stress, recent elasto-plastic with continuous yielding including such special properties as volume change, stress path, nonassociativeness, damage and softening, anisotropic hardening and viscoplasticity have been proposed and developed. Here, the main attention is devoted to the hierarchical single surface (HISS) concept developed by the authors and co-workers [1-6], in the context of the theory of plasticity.

Many of the above attributes are important for geologic materials (soils, rocks and concrete) and material contacts (interfaces and joints), which is the main concern herein. It may be noted that the models, analysis and results discussed here are applicable also to other classes of engineering materials and contacts such as in concrete, metals, composites and ceramics.

2. BRIEF BACKGROUND

The incremental finite element equations can be expressed as

$$[k]\{\Delta q\}=\{\Delta Q\}+\{\Delta Q_r\} \tag{1}$$

where [k] is the variable stiffness matrix, $\{\Delta q\}$, $\{\Delta Q\}$ and $\{\Delta Q_r\}$ are the incremental nodal displacement, applied load and residual load vectors, respectively. The stiffness matrix is given by

$$[k] = \int [B]^T [C][B] dV \qquad (2)$$

where [B] is the transformation matrix, [C] is the tangent constitutive or stress-strain matrix, and V is the volume. For different types of material responses, elastic, nonlinear elastic, elasto-plastic, viscoplastic, etc., the matrix [C] is defined by a set of material constants. For example, for linear elastic isotropic materials, they are Young's modulus E, and Poisson's ratio, ν or shear modulus G and bulk modulus K (or B), whereas for elasto-plastic materials, they will be E, ν and a_i ($i = 1, 2 \ldots n$), where the number of a_i will depend on various models such as perfectly plastic, hardening, softening, damage, and viscoplastic.

Here a plasticity based hierarchical single surface (HISS) modelling approach for solids and discontinuities is proposed by Desai and co-workers [1-6]. This approach, depicted in Fig. 1, allows for various types of behavior such as associative, nonassociative, anisotropic hardening, damage and softening, fluid pressure and viscoplastic. It also allows for such special features as effect of state of stress, stress paths, initial density, roughness at interfaces, volume change (dilation) and induced anisotropy.

The attention here is centered on the discussion of the influence of some of the above factors on the reliability of computer solutions and the need for special techniques required to deal with them. Before this discussion, a brief description of the modelling approach is given.

3. HIERARCHICAL MODELS FOR SOLIDS

For the basic δ_0-model following isotropic hardening and associative response, the yield surface F is expressed as

$$F = J_{2D}/p_a^2 - F_b\ F_s) \tag{3}$$

$$F_b = -\alpha(J_1/p_a)^n + \gamma(J_1/p_a)^2, \text{ and}$$

$$F_s = (1 - \beta S_r)^m$$

S_r = stress ratio $(\sqrt{27}/2)\ J_{3D}/J_{2D}^{1.5}$, p_a = atmospheric pressure, α, n, γ, β and m are response functions or parameters, J_i (i = 1, 2, 3) invariants of the stress tensor σ_{ij} and D denotes deviatoric. The basic function F_b is related to F in J_1 - J_{2D} space, Fig. 2(a), and the shape function F_s is related to F in the principal stress space, Fig. 2(c).

For the nonassociative case, the plastic potential function, Q, is expressed as the stem of F and the correction function $h(J_i, \xi)$:

$$Q = F + h\ (J_i,\ \xi) \tag{4}$$

where $\xi = (d\epsilon_{ij}^p\ d\epsilon_{ij}^p)^{1/2}$ and ϵ_{ij}^p = total incremental plastic strain tensor, and d denotes increment.

A simple form of hardening or growth function is given by

$$\alpha = a_1/\xi \tag{5a}$$

where a_1 and η_1 are hardening constants. Another growth function used to include effects of both volumetric and deviatoric components of plastic strains is given by

$$\alpha = \frac{h_1}{(\xi_v + h_3 \xi_D)^{h_2}} \tag{5b}$$

where h_i (i = 1, 2, 3) are material constants and ξ_v and ξ_D are the trajectories of volumetric and plastic strains, respectively.

Some of the special characteristics and advantages of the single continuous yield surfaces, F, Fig. 2, are (1) the function defines continuous yielding and involves no intersection of yield surfaces, (2) the final yield surface corresponds to the unique ultimate state, thus ambiguities due to the use of other definitions such as peak and failure are avoided, (3) the concept allows for easy incorporation of the change in shape and size of F, (4) the hardening function (α) depends upon both volumetric and deviatoric plastic strain trajectories, (5) different strengths in compression and extension are allowed for, (6) allows for initiation of volume dilation before peak stress, (7) the hierarchical approach allows for the devleopment of models of progressive complexities such as nonassociative (δ_1), anisotropic hardening with fluid pressure (δ_{2+p}), damage and softening (δ_{o+r}) and viscoplastic (δ_{o+vp}), and (8) the formulation possesses consistent theoretical basis.

Details of the above concept, verification of models and implementation for various static and dynamic problems are given elsewhere [1-8]. The list of the required material constants involved in the δ_o, δ_1, δ_{o+r} and δ_{2+p} models is given below:

Elasticity: E, ν or K and G (or B)

Plasticity:

Hardening: a_1, η_1, Eq. 5(a)

Ultimate: γ, β(F when α = 0)

Phase Change: n (Defines the state at which the volume change transits from compressive to dilative)

m = -0.5

κ = nonassociative parameter; defines h in Eq. (4).

r_u, k, R = damage parameters (in damage function $r = r_u (1 - k^R)$

h_1, h_2, h_3 . = hardening parameters (for cohesive soils), Eq. 5(b).

4. HIERARCHICAL MODEL FOR CONTACTS

The specialized form of F for planar contacts such as interfaces and joints is given by

$$F=(\tau/p_a)^2+\alpha(\sigma_n/p_a)^n-\gamma(\sigma_n/p_a)^2=0 \tag{6}$$

where τ and σ_n are the shear and normal stresses at the interface. The hierarchical formulation for contacts follows the same procedure as for solid with ξ given by

$$\xi=\int[(du_r^p)^2+(dv_r^p)]^{1/2} \tag{7}$$

where du_r^p and dv_r^p = increments of plastic shear and normal (relative) dislacements. The other aspects follow similar to those for solids; details are available elsehwere [6, 9].

5. DISCUSSION OF FACTORS

The discussion is presented under three sections (1) Material Parameters, (2) Special Characteristics, and (3) Implementation.

5.1 Material Parameters and Constants

Variability and sensitivity of material parameters are among the two factors that influence the reliability. Determination of the parameters from laboratory and/or field testing is an integral part of this consideration.

5.2 Variability

Most of the testing for finding material constants is performed in the laboratory. For geologic materials like soils and rocks, it is usually necessary to obtain specimens in the field and then bring them to the laboratory. This process itself can introduce considerable uncertainty and variability in the material parameters due to the sample disturbance, which is caused due to reasons such as change in the state of stress, and water content during the transportation from the field to the laboratory. This topic is indeed complex, and the effect of the sample disturbance on the material parameters is difficult to assess. Some empirical methods are available for limited number of constants involved in conventional models. However, they may become unreliable when a greater number of constants need consideration.

In contrast to specimens of metal and composites, where it is relatively easier to construct samples with the same (initial) properties, it is difficult to do so in the case of geologic materials. For instance, it is difficult to fabricate specimens with the same initial isotropy and density. As a consequence, specimens tested under the same loading conditions and path may exhibit different stress-strain-strength behavior. Hence, it is necessary to perform averaging, least square or optimization analysis, so as to obtain representative and weighted values of the constants. Use of optimization techniques is perhaps the most effective way; however, this area still needs continuing developments. At this time, the most common techniques used are based on the least square methods, which are also used for the results considered herein.

5.3 Sensitivity of Constants

Sensitivity of selected constants for the HISS models is discussed below. This is achieved in terms of the influence of the changes in these constants on the predictions from constitutive models themselves [10]. It is implied that when such models are implemented in computational procedures, the reliability of the computed results will be affected by the variation in the constants. In other words, for reliable and accurate results, it is necessary to evaluate the material constants as accurately as possible, based on appropriate laboratory and/or field tests coupled with consistent least square and/or optimization schemes.

5.3.1 Elastic Constants G and B (or K)

Values of elastic constants not only define the behavior of material inside the yield surface, it also can considerably affect the computed response during undrained virgin loading. Effect of the values of elastic constants, shear modulus (G) and bulk modulus (K or B), on the predicted undrained behavior of a typical normally consolidated clay is illustrated in Figs. 3 and 4. Figures 3(a) and (b) show shear stress vs. shear strain curves and the corresponding stress paths, respectively, predicted by the model with constant shear modulus (= 790 psi) and varying bulk moduli from 700 to 20,000 psi. Figures 4(a) and (b) illustrate the effect of G varying from 100 to 5000 psi on the predicted effective stress path and shear stress-shear strain curves, with constant B = 3667 psi. From Figs. 3 and 4, it can be seen that the undrained effective stress path is essentially independent of values of G, and strongly dependent on those of B. On the other hand, the initial slope and response in the predicted shear stress-shear strain curve is controlled by the values of G.

5.3.2 <u>Phase Change Parameter, n</u>

The phase change parameter n that defines the point at which the volume change transits from compression to dilation is among the factors that govern the shape of the yield surface, F, the equation for which can be expressed as

$$\left(\frac{J_{2D}}{J_{2D_m}}\right)\left(\frac{n-2}{n}\right)\left(\frac{2}{n}\right)^{\frac{2}{n-2}}-\left(\frac{J_1}{J_{1m}}\right)^2\left[1-\left(\frac{J_1}{J_{1m}}\right)^{n-2}\right]-0 \qquad (8)$$

where J_{1m} and J_{2Dm} are maximum values of J_1 and J_{2D} on the yield surface. From Eq. (8), it may be observed that the shape of the yield surface in the normalized stress space is controlled by the value of n. Yield functions for different values of n are plotted in Fig. 5. It is observed here that the yield surface becomes skewed with increasing n values, and at the limit (n→∞) the yield surface tends to the triangular shape.

5.3.3 <u>Ultimate Parameter, β</u>

Shape of the yield surface on the octahedral plane is governed by the value of β. Here study of traces of yield surfaces on octahedral plane for different values of β showed that the yield surface is circular in the octahedral plane for $\beta = 0$, and with increasing values of β the shape of the yield surface tends to a triangular shape with rounded corners. When $\beta > 0.77$, yield surface becomes non convex. Figure 6 shows the shape of yield surfaces in the p-q space where $q = \sigma_1 - \sigma_3$, and $p = (\sigma_1 + \sigma_2 + \sigma_3)/3$, for different values of β. It can be observed from this figure that the relative size between the compression and extension regions of the yield surface is controlled by the value of β. Higher the β value, higher the difference between compression and extension strengths of the soil.

5.3.4 Hardening Parameter, h_3

This parameter defines the effect of plastic shear strain on hardening. Figure 7 shows predicted effective stress paths for different values of h_3; $h_3 = 0$ indicates that hardening is controlled only by the volumetric plastic strains, while the influence of the deviatoric strain increases with increasing values of h_3.

5.3.5 Nonassociative Parameter, κ

Nonassociative parameter, κ, in the δ_1 model controls the volumetric behavior of (drained) shear tests. Figure 8 shows the effect of on the predicted volumetric response of a typical granular material. It is observed here that the variation of α from 0 to 1 changes the volumetric response that ranges from compactive to dilative. However, this variation had only small effect on the shear stress-shear strain response.

6. THIN LAYER ELEMENT FOR CONTACTS

The thin layer element concept [11] is used here to simulate the interface or joint as a "smeared" zone of finite thickness (t) with properties different from those of the surrounding elements. For the finite element calculations herein, the ratio of thickness t to the width B (t/B) of the element is 0.01, Fig. 9.

The influence of κ for the contact or joint (in concrete) is shown in Figure 10 for a typical rough joint (with asperity angle of 5 degrees), [6, 9]. From the finite element analysis, it can be seen that for the joint considered, the computed results compare well with laboratory test data [6, 9] for the value of κ equal to 0.6 to 0.7.

The above analysis indicates that the sensitivity of predictions and computer solutions to variations in various constants can be expressed in the relative order of decreasing importance as hardening and softening, phase change, elastic and ultimate.

7. IMPLEMENTATION

In a nonlinear iterative finite element procedure, the computer module representing the constitutive relation

$$\{d\sigma\}-[C^{ep}]\{d\epsilon\} \tag{9}$$

where $\{d\sigma\}$ and $\{d\epsilon\}$ are vectors of incremental stress and strain, respectively, and $[C^{ep}]$ is the tangent elasto-plastic matrix, may be called thousands of times. Also, with an advanced plasticity based model for anisotropic hardening response, it is very important to evolve an efficient and robust scheme to affect the drift correction procedure in order to satisfy the consistency condition during the incremental stress or strain and the iterations therein. Such a drift correction procedure is developed [10] by taking into account the best properties in available procedures such as the subincrement method [12], drift correction method [13], and the elastic predictor plastic correction method [14].

7.1 Basic Formulation of Drift Correction Scheme

Figure 11 schematically shows the iterative procedure used during virgin loading conditions. The problem may be expressed as

Given σ_{ij}^{o}, α^{o}, ξ_i^{o}, $d\epsilon_{ij}^{t}$ and

$$F\ (\sigma_{ij}^{o},\ \alpha^{o}) = 0 \tag{10}$$

Find σ_{ij}^{c}, α^{c}, ξ_i^{c}

which satisfy following relationships

$$F\ (\sigma_{ij}^{c},\ \alpha^{c}) = 0 \tag{11}$$

$$d\varepsilon_{ij}^{t} = d\varepsilon_{ij}^{e} + d\varepsilon_{ij}^{p} \tag{12}$$

$$\sigma_{ij}^{c} = \sigma_{ij}^{o} + C_{ijkl}^{e}\ d\varepsilon_{kl}^{e} \tag{13}$$

$$\xi_i^{c} = \xi_i^{o} + f_i\ (d\varepsilon_{ij}^{p}) \tag{14}$$

$$\alpha^{c} = f_{\alpha}\ (\xi_i^{c}) \tag{15}$$

Where superscripts o and c refer to quantities associated with the state O and C in Fig. 11, $d\epsilon_{ij}^{t}$, $d\epsilon_{ij}^{e}$ and $d\epsilon_{ij}^{p}$ are total, elastic and plastic incremental strain tensors, respectively. C_{ijkl}^{e} is the elastic constitutive tensor, ξ_i are different trajectories of plastic strains such as ξ, ξ_D and ξ_v. The functions f_i relate incremental plastic strains to incremental strain trajectories, and f_{α} is the hardening function.

Equations (10) to (15) are solved in two stages using a predictor corrector algorithm. The intermediate stage I is found from

$$\sigma_{ij}^{I} = \sigma_{ij}^{o} + C_{ijkl} d\epsilon_{k}^{t} \tag{16}$$

Elastic predictor plastic corrector method [14] uses elastic constitutive tensor, C_{ijkl}^{e} in Eq. (16), and plastic predictor plastic corrector method (drift correction method [13]) uses the initial elasto-plastic matrix, C_{ijkl}^{ep} in Eq. (16). It was found that both methods lead to similar results for the case of small strain increments. However, for large strain increments, special procedures may be necessary to handle some difficulties which are described elsewhere [10].

After algebraic manipulations, σ_i^c can be obtained as

$$\sigma_{ij}^{c} = \sigma_{ij}^{I} + \frac{F\,(\sigma_{ij}^{I}, \alpha^{I})\; C_{ijkl}^{e}\; n_{kl}^{Q}}{\frac{\partial F}{\partial \alpha}\Big|_{I} f_i\,(n_{ij}^{Q}) - \frac{\partial F}{\partial \sigma_{mn}}\Big|_{I} C_{mnop}^{c}\; n_{op}^{Q}} \qquad (17)$$

where

$$n_{ij}^{Q} = \frac{\partial Q}{\partial \sigma_{ij}} / [(\frac{\partial Q}{\partial \sigma_{mn}} \frac{\partial Q}{\partial \sigma_{mn}})^{1}/2] \qquad (18)$$

and Q is the potential function.

Special procedures were developed for cases where (1) $J_1^I < 0$, (2) $J_1^c < 0$, (3) $F\ (\sigma_i^I, \alpha^I) < 0$, and (4) $F\ (\sigma_{ij}^c, \alpha^c) < 0$.

Due to space limitations, details of these procedures are not given here but can be found in [10].

7.2 Effect of Subincrements on Iterative Algorithm

The effect of magnitudes of subinbcrements on the accuracy and reliability of the constitutive model module is studied using the proposed plastic predictor plastic corrector (drift correction) method. First, the accurate strain path for a TC test (triaxial compression with J_1 = constant) [15] is calculated by using the stress to strain algorithm [10]. Then two points in this strain path, one before the phase change line and one after the phase change line, were selected for numerical experiments.

Stress paths for three strain increments of the order of

10^{-3} are predicted using three sizes of subincrements. Figure 12(a) shows the predicted stress path including iteration steps for subincrements of the order of 10^{-3}, 10^{-4} and 10^{-5} in the compaction region. Note that the first coincides with no subincrements. Figure 12(b) shows the similar plots for the numerical experiments in the dilatation region. Following observations may be made from these figures: (1) all the strain increments converged in two to three iterations, (2) predicted stress paths are 'accurate' only for smaller subincrements, (3) return path is not perpendicular to the potential surface, and can make very small angles with it, (4) convergence does not necessarily imply that the final answer is correct; therefore, it is important to limit the maximum size of a strain increment, (5) since the return path can make small angles with the yield surface, the exact stress point where $\partial Q/\partial \sigma_{ij}$ is calculated is very important. With the elastic predictor plastic corretor method, it was found that if $\partial Q/\partial \sigma_{ij}$ is calculated at I (Fig. 11) as suggested in [14], the solution drifted away, especially in the dialation region, and (6) from the curved nature of the predicted stress path in Figs. 12(a) and (b), it is concluded that the plastic prediction is very sensitive to the stress point where $C^{ep}_{ijk\partial}$is calculated in the plastic predictor plastic corrector method. For the numerical experiments conducted for a TC test, it was found that if $C^{ep}_{ijk\partial}$ is calculated at $[\sigma_{ij}^{o} + 0.35\ (\sigma_{ij}^{I} - \sigma_{ij}^{o})]$, improved results are obtained. Since σ_{ij}^{I} is not known in advance, it is necessary to perform few iterations here.

7.3 Application: Dynamics of Porous Media

The anisotropic hardening (δ_{2+p}) constitutive model with the above drift correction scheme has been implemented in a nonlinear dynamic finite element procedure based on the generalized Biot's theory for dynamics of porous media. This program was used to back predict the field response of an instrumented pile segment under cyclic loading [10, 16]. The finite element mesh for the pile segment subjected to cyclic loading is shown in Fig. 13. Here various stages such as initial stress, driving of the pile, consolidation and cyclic loading are simulated as they occurred in the field. Typical comparisons between predictions and observations for shear stress vs. relative displacement of pile are shown in Fig. 14. The correlation between the two is considered to be highly satisfactory.

8. DAMAGE AND SOFTENING

The damage model, δ_{0+r}, among the hierarchical models, considers strain softening as the performance of the structure, with a damage parameter r, which represents the ratio between the volume of the damaged or fractured part, and the total volume [3, 4, 17]. The damage parameter, r, relates the mean stress to the topical or continuum stress in the undamaged part of the material as follows:

$$\sigma^{m}_{ij}-(1-r)\sigma'_{ij}-\frac{r}{3}\delta_{ij}\sigma'_{kk} \qquad (19)$$

Here r is related to the deviatoric plastic strain trajectory $_{D}$ as follows:

$$r = r_u \ (1-\exp(-k(\xi_D)^R)) \qquad (20)$$

in which r_u is thc ultimate value or r, and k, R are damage constants.

Introducing of the damage model in the finite element procedure leads to the following equation:

$$\int_v[B]^T[L][B]dV\{dq\} = \{Q]-\int_v[B]^T\{\sigma^m\}dV+\int_v[B]^T\{S^t\}drdV \qquad (21)$$

in which {dq} is the vector of incremental displacement, {Q} is the vector of external loading, $\{\sigma^m\}$ is the vector of mean stress, $\{S^t\}$ is the vector of deviatoric topical stress, dr is the increment of the damage parameter for the last iteration and [L] is the nonsymmetric constitutive tensor expressed as

$$L_{ijk\ell} = (1-r)\, C^{ep}_{ijk\ell} + \frac{r}{3}\,\delta_{ij}\, C^{ep}_{mmk\ell} \qquad (22)$$

Alternatively, we can leave the symmetric $[C^{ep}]$ on the left-hand-side, and transfer a term corresponding to the non-symmetric component to the right side as

$$\int_v[B]^T[C^{ep}][B]dV\ \{dq\} = Q - \int_v[B]^T\{\sigma^m\}dV$$
$$+\int_v[B]^T\{S^t\}drdV + \int_v[B]^T\{dS\}^t\}rdV \qquad (23)$$

in which $\{dS^t\}$ is the vector of the increment of the deviatoric topical stress for the last step. Both these strategies can help to overcome the numerical difficulties by avoiding an ill-conditioned matrix on the left-hand-side when damage and subsequent softening occurs.

8.1 Application

The damage model with the foregoing drift correction scheme is implemented in a nonlinear finite element procedure for both static and dynamic analysis [17]. Typical examples are given below.

Figure 15 shows the mesh for the quarter of a concrete block subjected to uniform compression q on the side; meshes with 1-, 4- and 16-elements were used. Uniform displacements are prescribed on the top of the block. The material constants for concrete are given below.

$E = 37000\ N/mm^2$; $\nu = 0.25$; $N = 5.237$; $\gamma = 0.06784$; $\beta = 0.7553$; $m = -0.5$; $a_1 = 4.714E-11$; $\eta_1 = 0.8262$; $r_u = 0.0, 0.2, 0.4, 0.6$; $k = 968.8$; and $R = 1.503$.

The results in terms of stress and strain (at the lower left Gauss point) for the static analysis are shown in Fig. 16 for four different values of r_u. The proposed procedure provided consistent results for the damage problem.

8.2 Time Integration and Damage Problem

For dynamic nonlinear problems involving such complexities as damage and softening, it is also necessary to use appropriate time integration scheme. Here, the the Generalized Time Finite Element Method (GTFEM) proposed in [18, 19] is used. It involves a weighted residual approach and considers the equilibrium over a time period instead of the equilibrium at a certain time level. The vector of weighting functions {W} is given as

$$\{W\} = \begin{Bmatrix} w_o \\ w_1 \end{Bmatrix} = \begin{bmatrix} 1-\tau & (1-\tau_2)^2 & (1-\tau_3)^3 & (1-\tau_4)^4 \\ \tau & \tau^2 & \tau^3 & \tau^4 \end{bmatrix} \begin{Bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{Bmatrix} \qquad (24a)$$

In which α_n are weighting parameters, and

$$\tau = \frac{t-t_{i-1}}{t_i - t_{i-1}} \qquad (24b)$$

Applying the weighting functions onto the finite element equations and integrating, and, at the same time, using the relation

$$\begin{Bmatrix} H_1 \\ H_2 \\ H_3 \\ H_4 \end{Bmatrix} = \sum_{n=1}^{4} \alpha_n \begin{Bmatrix} 1 \\ 1/(n+1). \\ 1/(n+1)(n+2) \\ 2/n+2 \end{Bmatrix} \qquad (25)$$

we arrive at the time integration recurrence formula

$$(H_1\frac{1}{\Delta t^2}[M] + H_2\frac{1}{\Delta t}[C] + H_3[K]\,d_{i+1} \tag{26}$$
$$= F_i + (2H_1\frac{1}{\Delta t^2}[M] - H_4[K])d_i - (H_1\frac{1}{\Delta t^2}[M] - H_2\frac{1}{\Delta t}[C] + H_3[K])d_{i-1}$$

For linear problems, [K] matrix is constant, and the above method yields the same results as the Newmark method. However, for nonlinear problems, [K] is not evaluated at specific time step, but is evaluated at a displacement value d_θ, which is weighted between d_{i-1}, d_i and d_{i+1}, Fig. 17:

$$d = (1/2 + \theta_1 - \theta_2)d_{i-1} + (1/2 - 2\theta_1 + \theta_2)d_i + \theta_1 d_{i+1} \tag{27}$$

The weighting values of θ_1 and θ_2 can be adjusted to improve accuracy [18, 19].

Since the equilibrium is averaged over a time period, the GTFEM is found to be more stable than the Newmark method when treating problems with highly nonlinear material properties such as that involving damage and softening. One-dimensional problems with the plastic Ramberg-Osgood model, and two-dimensional problems with the δ_{o+r} damage model were solved, and it was shown that when using the unconditionally stable Newmark method ($\gamma = 0.5$, $\beta = 0.25$), the results became unreliable when the time step approached the natural period of the system. Whereas, using the GTFEM, the results were still stable [17].

With the damage model, the concrete block, Fig. 18, is subjected to a dynamic load on the top, Fig. 19. The material constants are the same as in the static problem, except that r_u is taken as 0.5. When the time step is taken as 0.001 second, both the Newmark method and GTFEM gave exactly the same solution, which is considered as the "accurate" one (Fig. 20a). It can be seen that the natural period of the system is approximately 0.08 second. When the time step is increased, the solutions of the two methods start to differ. When the time step reaches 0.05 seconds, which is comparable with the period, the solution from the Newmark method starts to diverge after about 1.5 seconds (Fig. 20b), and when the time step reaches 0.1, which is larger than the period, it diverges after 0.5 seconds (Fig. 20c). However, in both cases, GTFEM gave stable solutions.

Thus, in addition to proper attention to the plasticity based constitutive model and associated drift correction scheme, it is also necessary to use an appropriate time integration scheme for nonlinear dynamic problems.

9. CONCLUSIONS

The study presented herein shows that (1) accurate determination of material constants in constitutive models is important for reliable computer solutions because a small change in values of some of them can cause significant changes in the computed solutions, (2) for highly nonlinear static and dynamic problems involving anisotropic hardening, and damaged and softening materials, it is necessary to develop and use proper drift correction and time integration schemes; otherwise, the computed solutions may lose reliability and robustness.

ACKNOWLEDGMENTS

A part of the research herein was supported from Grant Nos. CEE 8215344, CEE 8320256, ECE 8415417, and MSM 8618914 from the National Science Foundation, Washington, D.C., and Grant No. AFOSR 830256 from the Air Force Office of Scientific Research, Bolling AFB, Washington, D.C.

REFERENCES

[1] C.S. Desai, A general basis for yield, failure and potential functions in plasticity, Int. J. Num. Analyt. Meth. Geomech. (1980) 361-375.

[2] C.S. Desai and M.O. Faruque, Constitutive model for geological materials, J. Eng. Mech. Div., ASCE 110 (1984) 1391-1408.

[3] C.S. Desai, S. Somasundaram and G. Frantziskonis, A hierarchical approach for constitutive modelling of geologic materials, Int. J. Num. Analyt. Meth. Geomech. 10 (1986) 225-257.

[4] G. Frantziskonis, and C.S. Desai, Constitutive model with strain softening, Int. J. Solids Struct. 23 (1987) 733-750.

[5] S. Somasundaram, and C.S. Desai, Modelling and Testing for Anisotropic Behavior of Soils, J. Eng. Mech. Div., ASCE 114 (1988), 1473-1496.

[6] C.S. Desai and K.L. Fishman, Plasticity based constitutive model with associated testing for joints, Int. J. Rock Mech. Min. Sc. (1988) under publication.

[7] C.S. Desai and H.M. Galagoda, Earthquake analysis with generalized plasticity model for saturated soils, J. Earthquake Eng. Struct. Dyn. 18 (1989) 903-919.

[8] C.S. Desai and Q.S.E. Hashmi, Analysis, evaluation and implementation of a nonassociative model for geologic materials, Int. J. Plasticity 5 (1989) 397-420.

[9] K.L. Fishman and C.S. Desai, A constitutive model for hardening behavior of rock joints, Proc. 2nd Int. Conf. Constitutive Laws Eng. Mat., Elsevier, New York (1987).

[10] G.W. Wathugala, Finite element dynamic analysis of nonlinear porous media with applications to piles in saturated clays, Ph.D. Dissertation, Dept. of Civil Eng. & Eng. Mech., Univ. of Arizona, Tucson, AZ (1990), under preparation.

[11] C.S. Desai, M.M. Zaman, J.G. Lightner and H.J. Siriwardane, Thin-layer element for interfaces and joints, Int. J. Num. Analyt. Meth. Geomech. 8 (1984) 19-43.

[12] M.O. Faruque and C.S. Desai, Implementation of a general constitutive model for geologic materials, Int. J. Num. Analyt. Meth. Geomech. 9 (1985) 415-436.

[13] D.M. Potts and A. Gens, A critical assessment of methods of corrections for drift from the yield surface in elasto-plastic finite element analysis, Int. J. Num. Analyt. Meth. Geomech. 9 (1985) 149-159.

[14] M. Ortiz and J.C. Simo, An analysis of a new class of integration algorithms for elasto-plastic constitutive relations, Int. J. for Num. Meth. in Engg., 23 (1986) 353-366.

[15] C.S. Desai and H.J. Siriwardane, Constitutive laws for engineering materials, Prentice-Hall, Englewood Cliffs, N.J., USA (1984).

[16] Earth Technology Corporation, Pile segment tests - Sabine pass, some aspects of the fundamental behavior of axially loaded piles in clay soils, ETC Report 85-007, Long Beach, CA (1986).

[17] L. Woo, Dynamic nonlinear finite element analysis for damage and softening materials, Ph.D. Dissertation, Department of Civil Eng. & Eng. Mech., Univ. of Arizona, Tucson, AZ (1990) under preparation.

[18] J. Kujawski and C.S. Desai, Generalized time finite element algorithm for nonlinear dynamic problems, Eng. Comp. 1 (1984) 247-251.

[19] C.S. Desai, J. Kujawski, C. Miedzialowski and N. Ryzynski, Improved time integration of nonlinear dynamic problems, Comp. Meth. Appl. Mech. Eng. 62 (1987) 155-168.

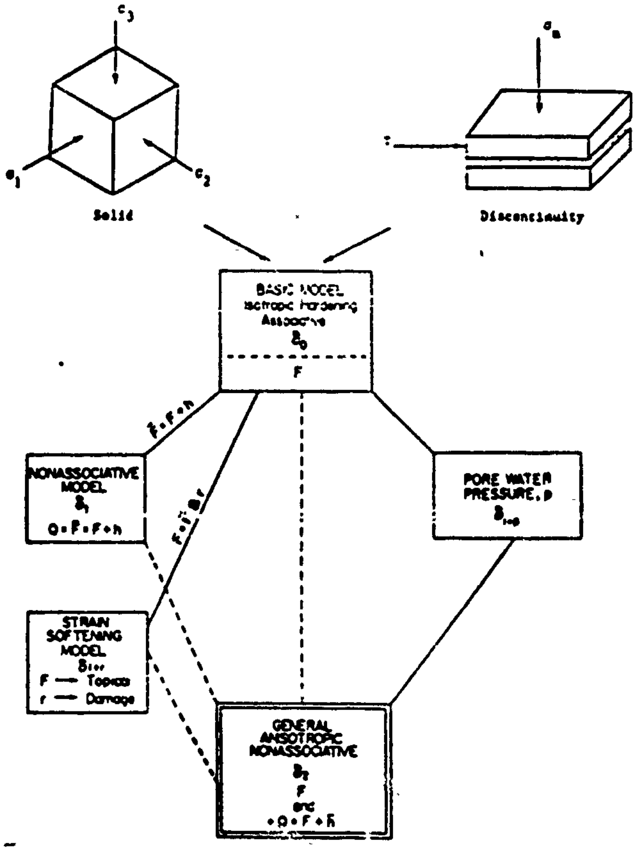


Fig. 1. Hierarchical single surface models

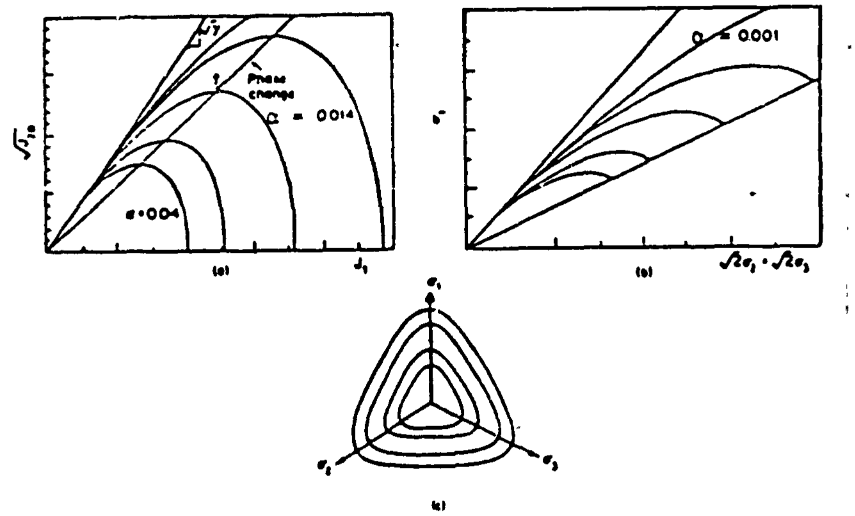
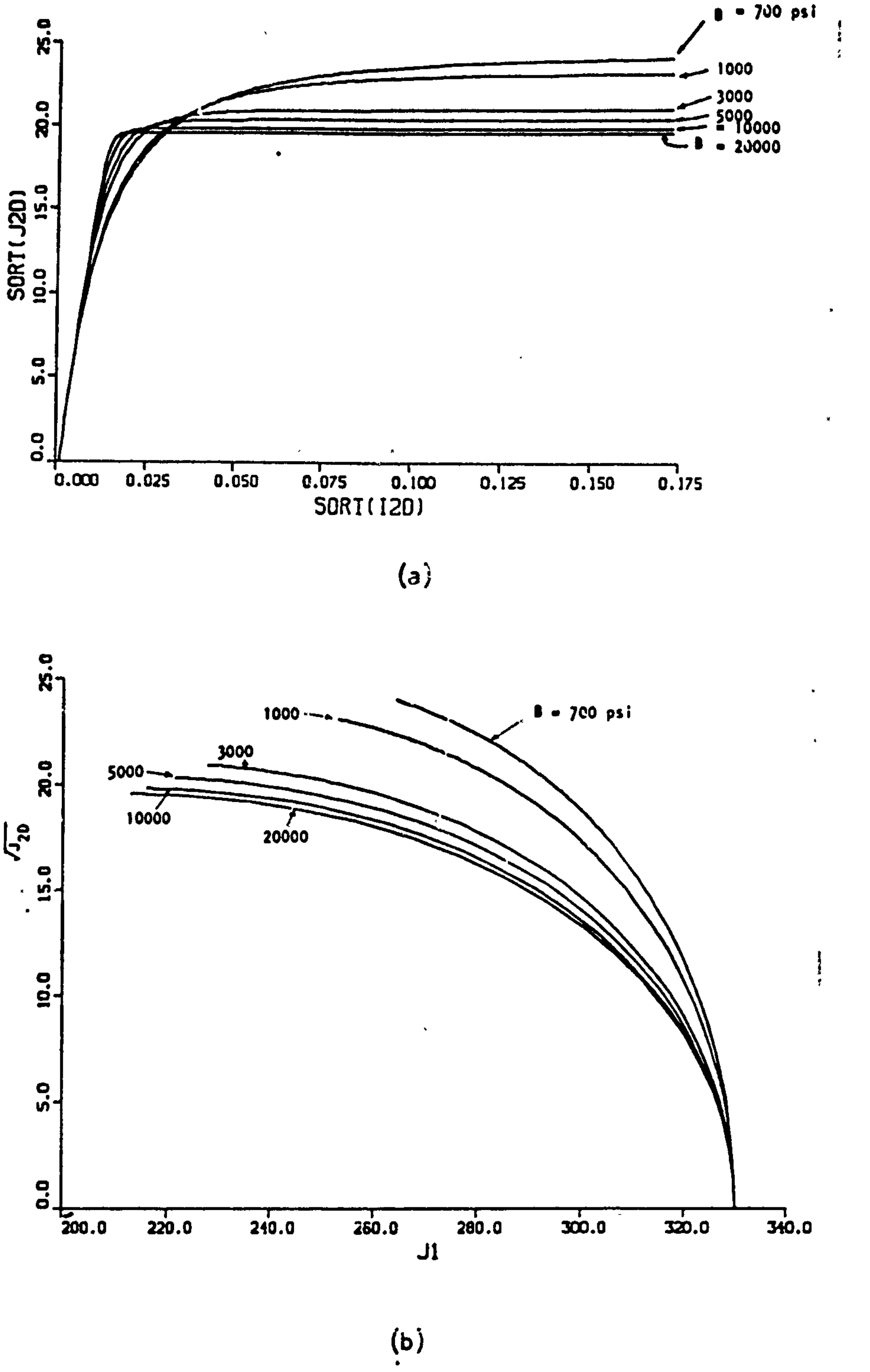


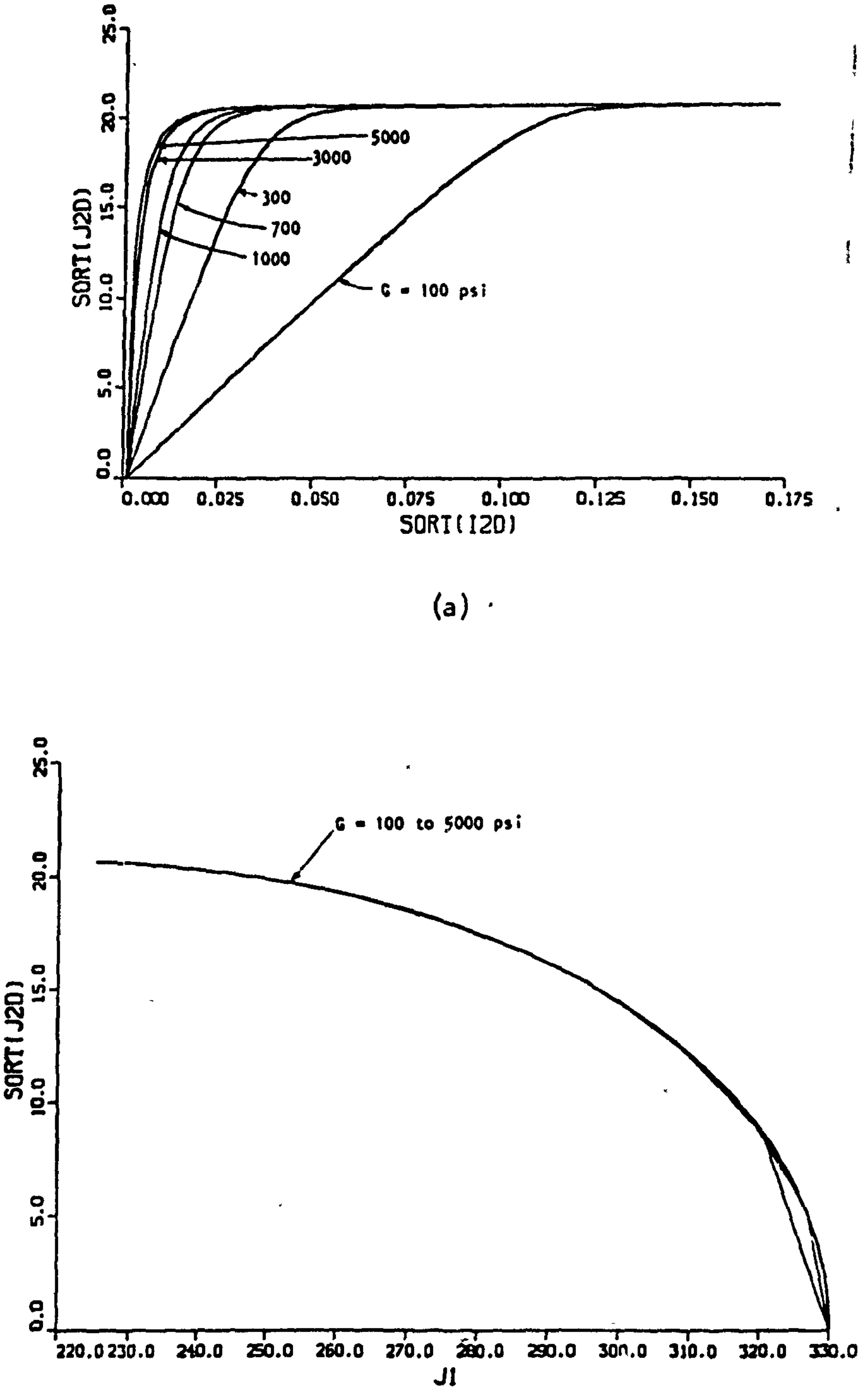
Fig. 2. Plots of F in various stress spaces



(a)

(b)

Fig. 3. Effect of varying bulk modulus on predicted (a) stress strain and (b) stress path responses; I_{2D} = second invariant of deviatoric strain tensor



(a)

(b)

Fig. 4. Effect of varying shear modulus on predicted (a) stress-strain, and (b) stress path responses

Effect of n

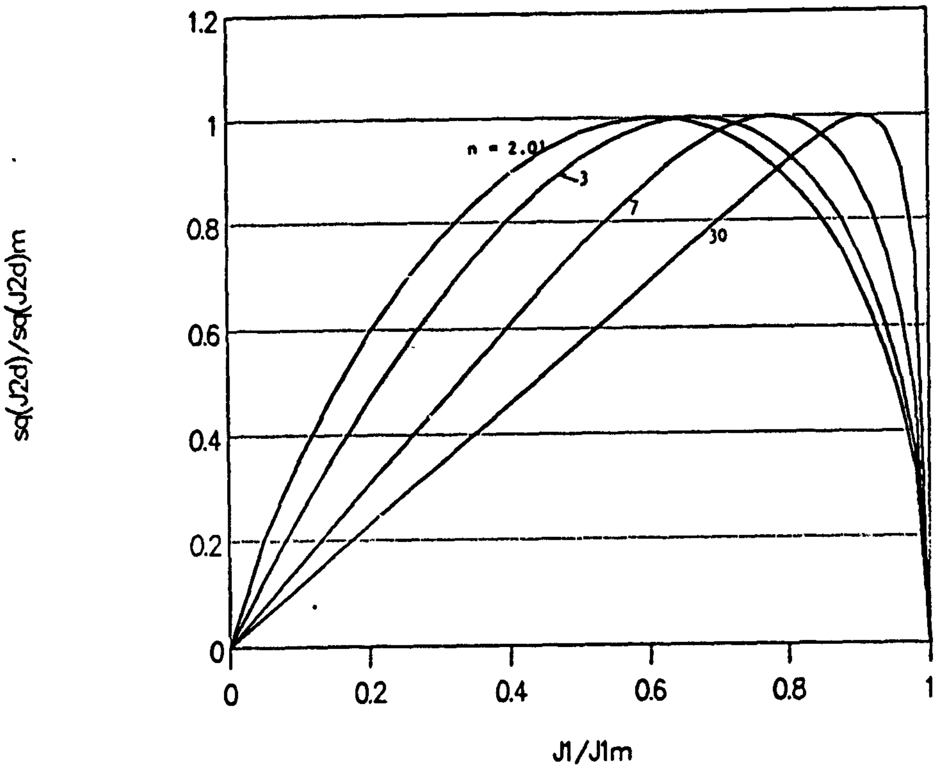


Fig. 5. Effect of phase change.parameter n on shape of F

Effect of Beta on the Shape of Yiled Surface

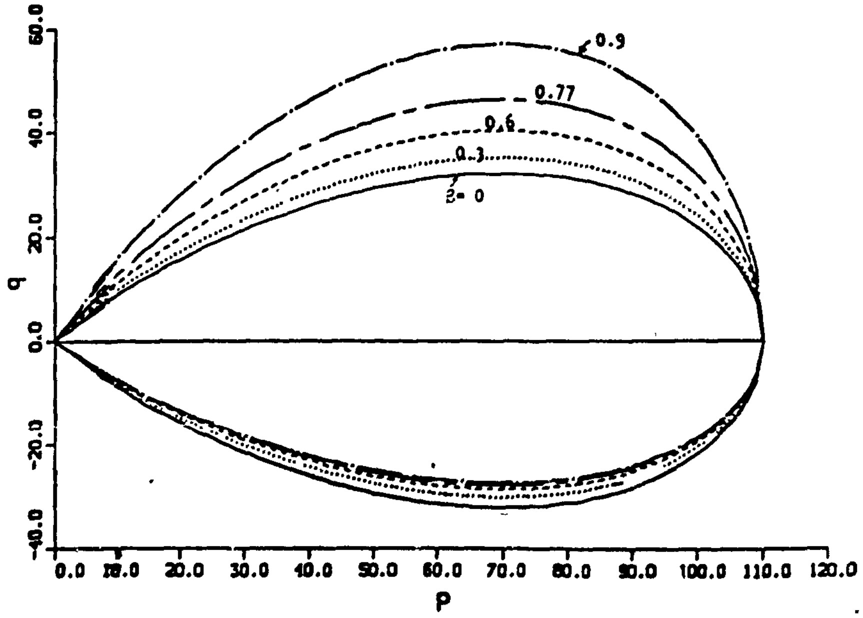


Fig. 6. Effect of β on shape of F

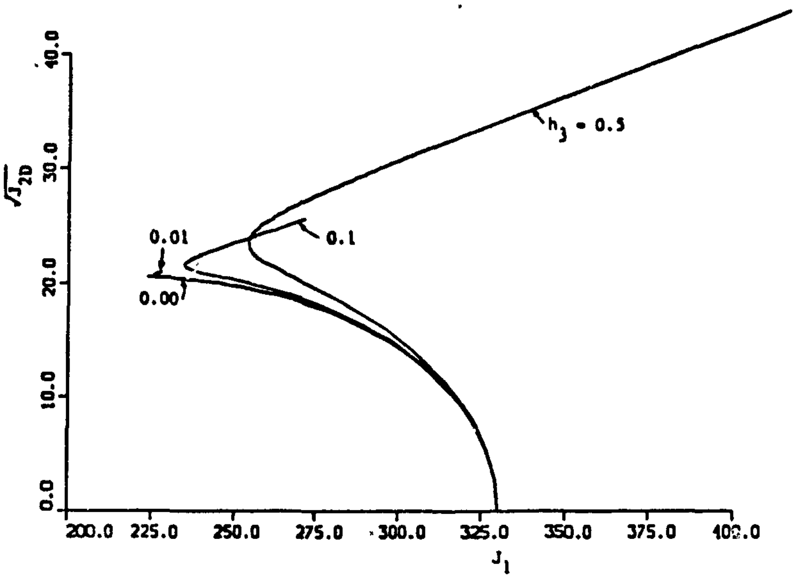


Fig. 7. Effect of h_3, Eq. (5b), on predicted stress paths

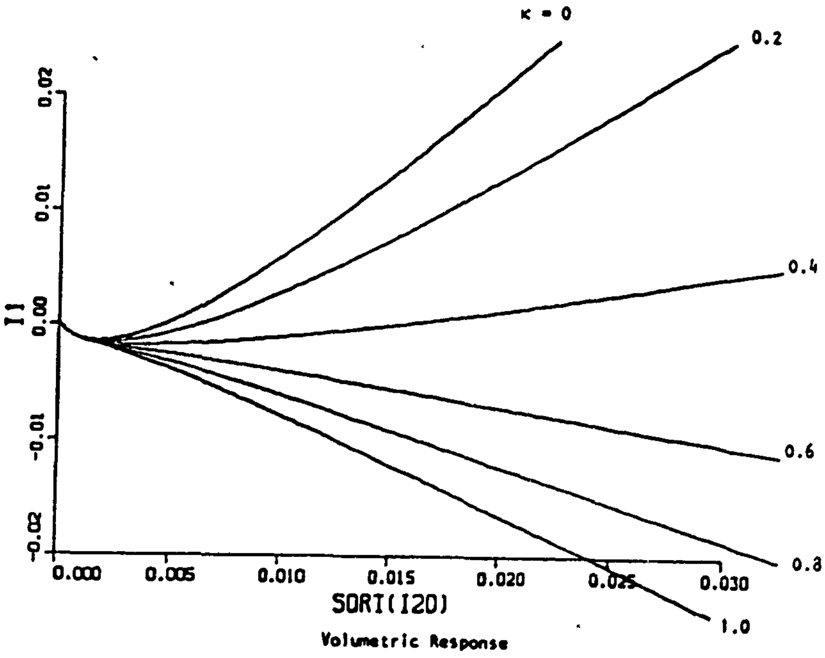
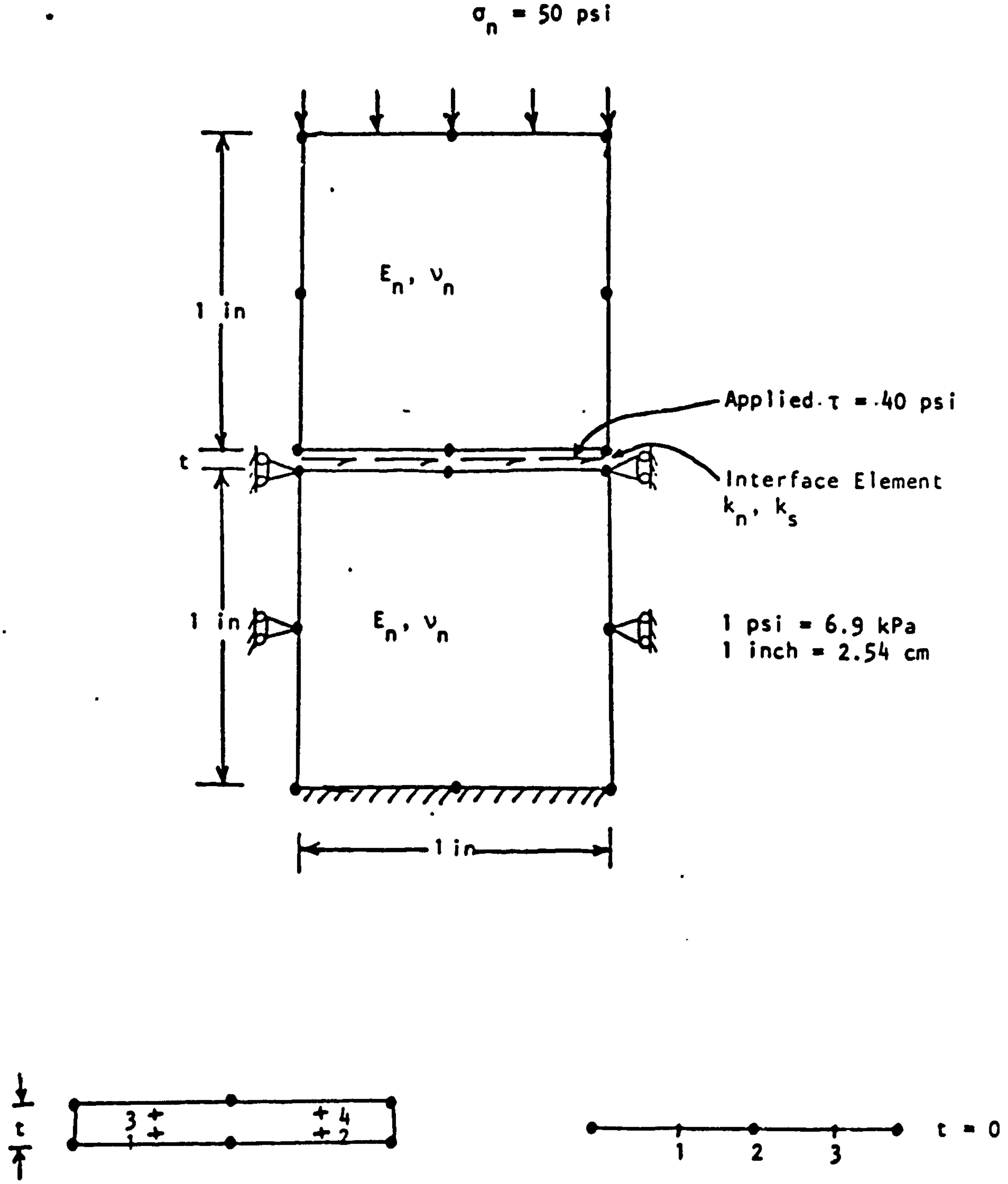


Fig. 8. Effect of nonassociative parameter κ on volume change behavior;$I_1 = \varepsilon_1+\varepsilon_2+\varepsilon_3$



Thin-layer: 4 Gauss points

Zero thickness: 3 Gauss points

Fig. 9. Thin layer element and example problem

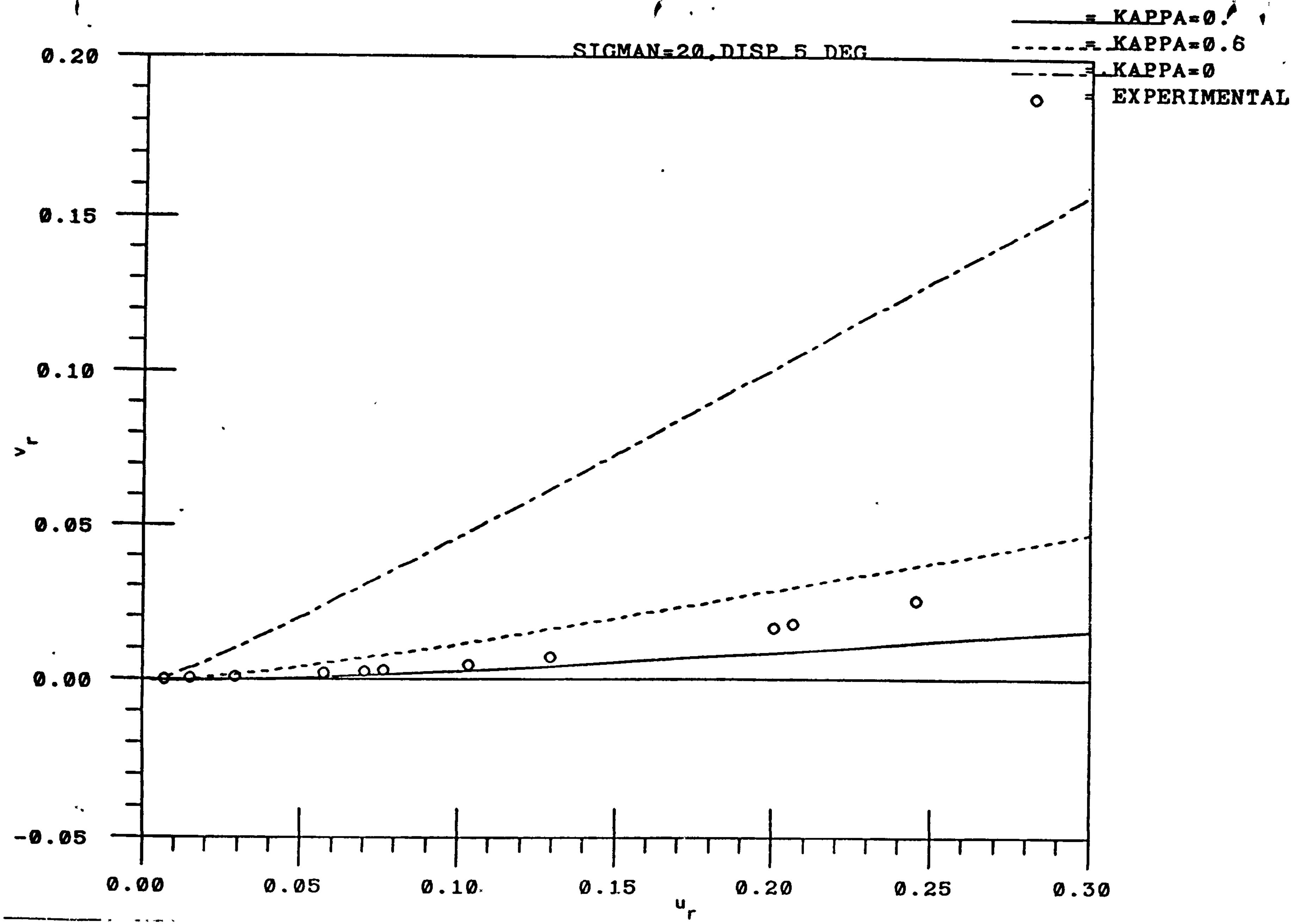


Fig. 10. Effect of nonassociative parameter on dilative behavior; u_r, v_r = relative shear and normal displacements, respectively

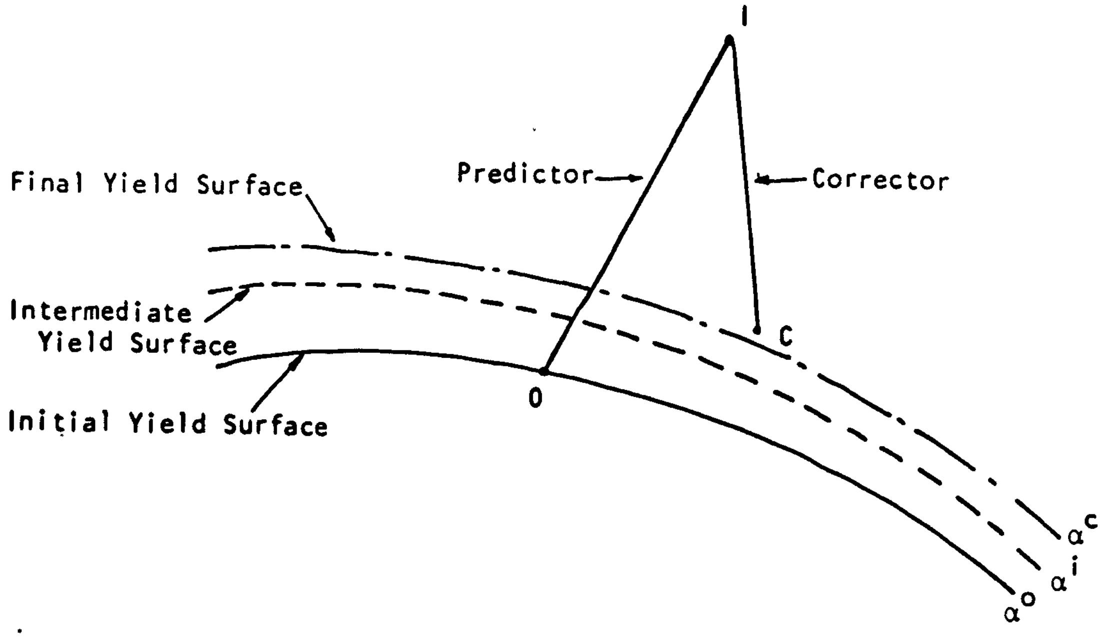


Fig. 11. Schematic of drift correction procedure

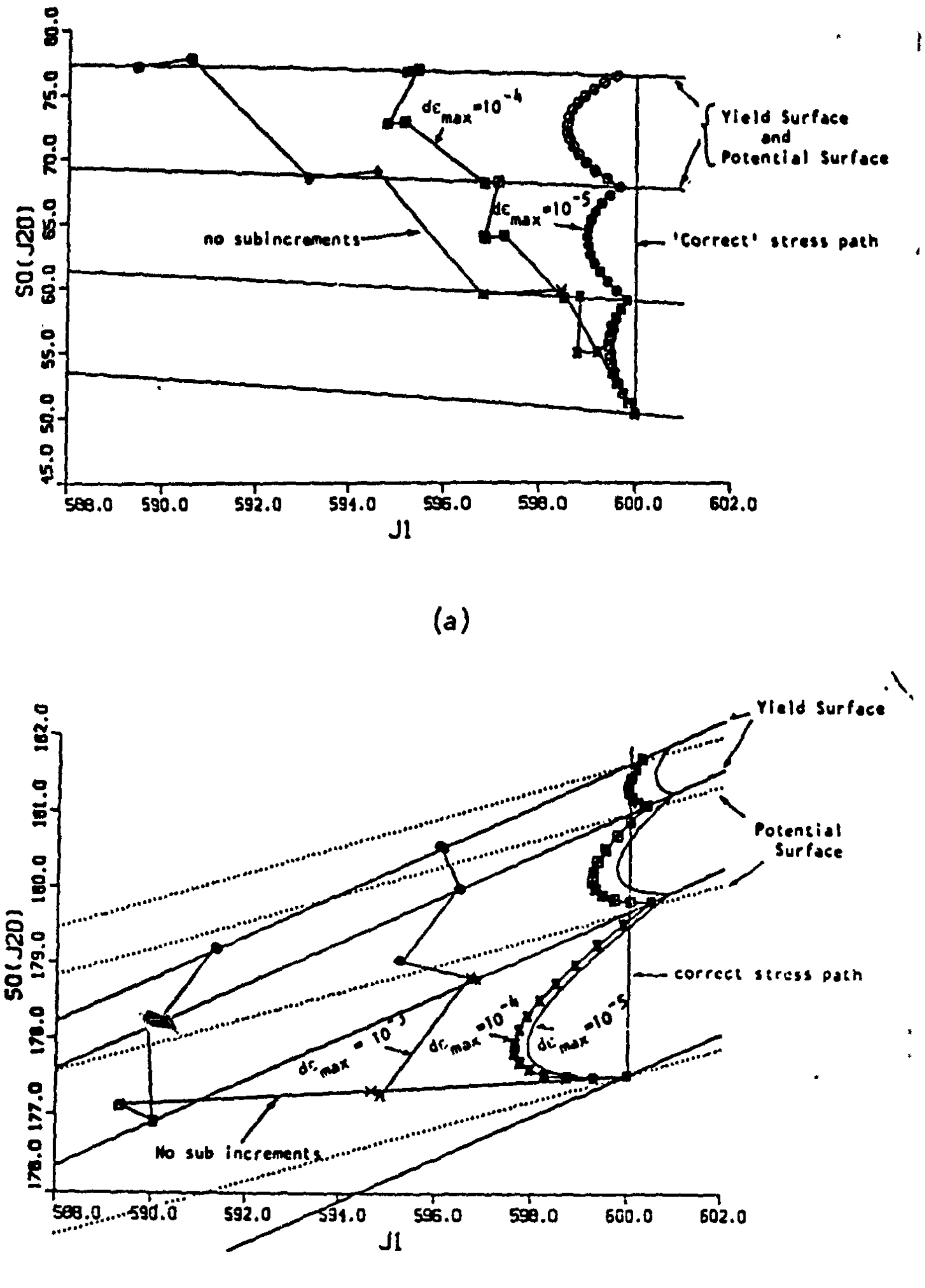


Fig. 12. Effect of strain increment size on convergence
(a) in compaction region, (b) in dilatation region

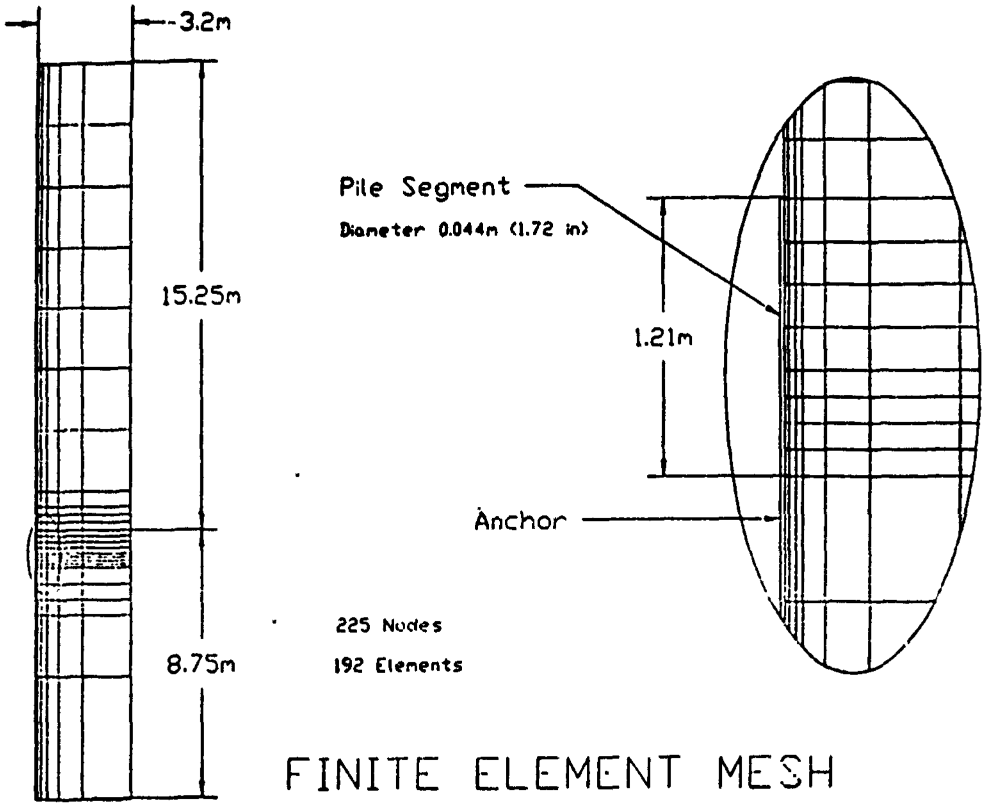


Fig. 13. Finite element mesh for cyclically loaded pile

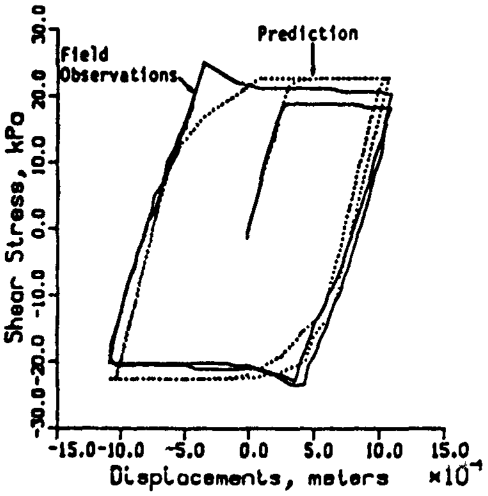


Fig. 14. Comparison of predicted shear stress vs. relative displacements for pile probiem

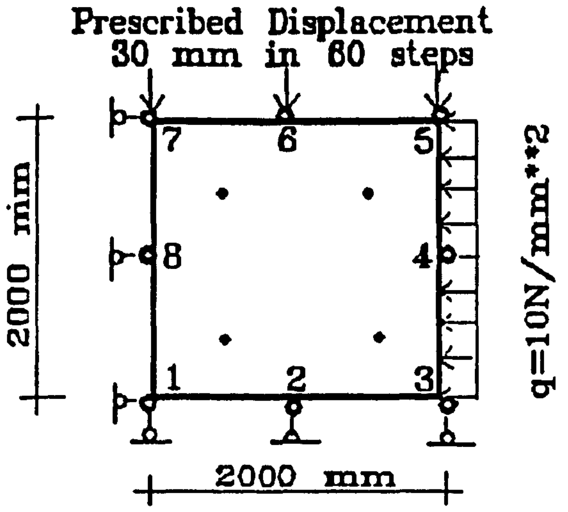


Fig. 15. Mesh for static problem in damage and strain softening concrete

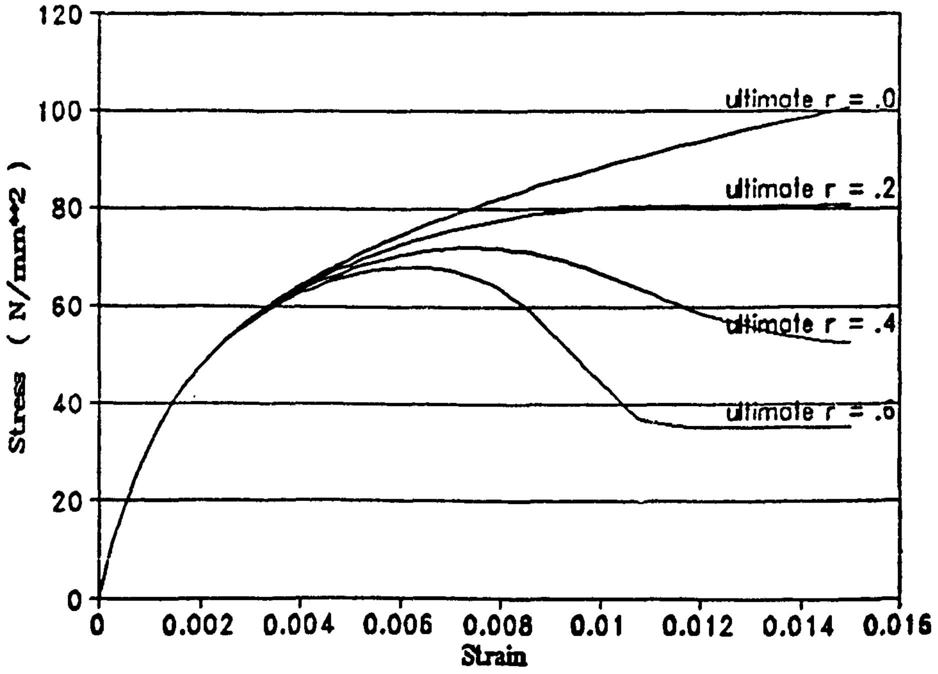


Fig. 16. Predicted softening static stress-strain response

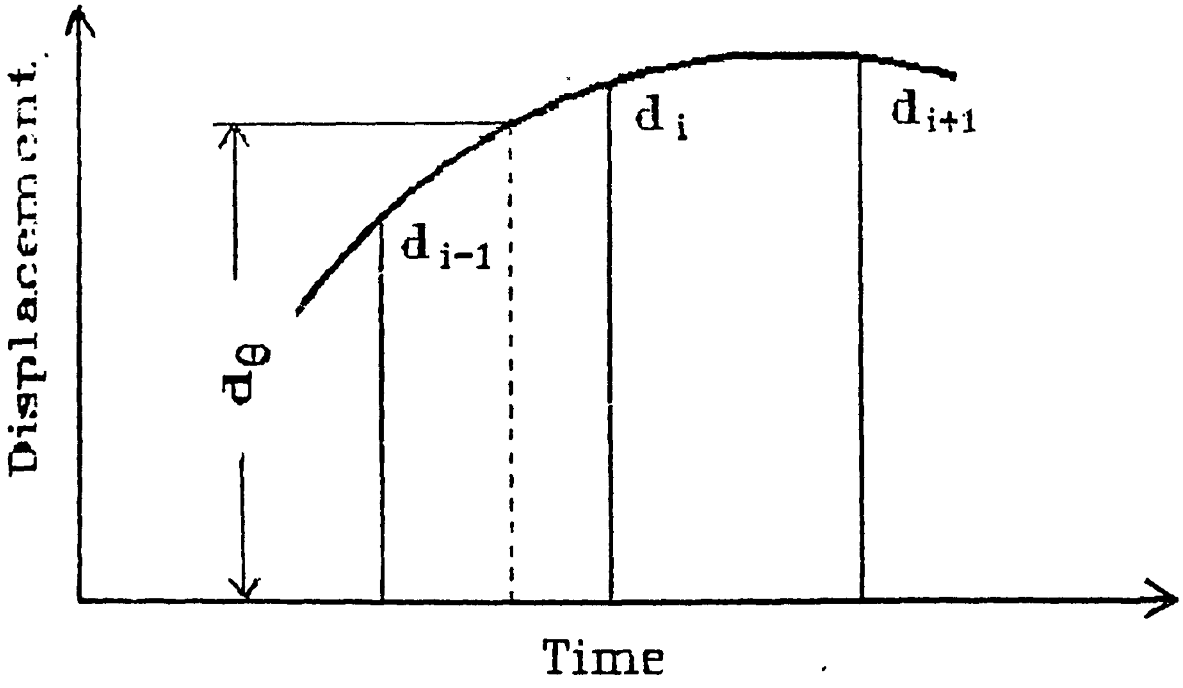


Fig. 17. Weighting for d_θ in GTFEM scheme

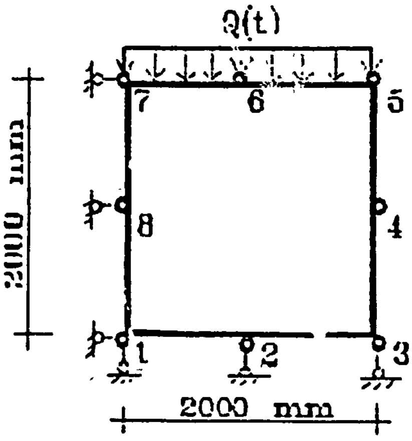


Fig. 18. Mesh for dynamic problem in damage and strain softening concrete

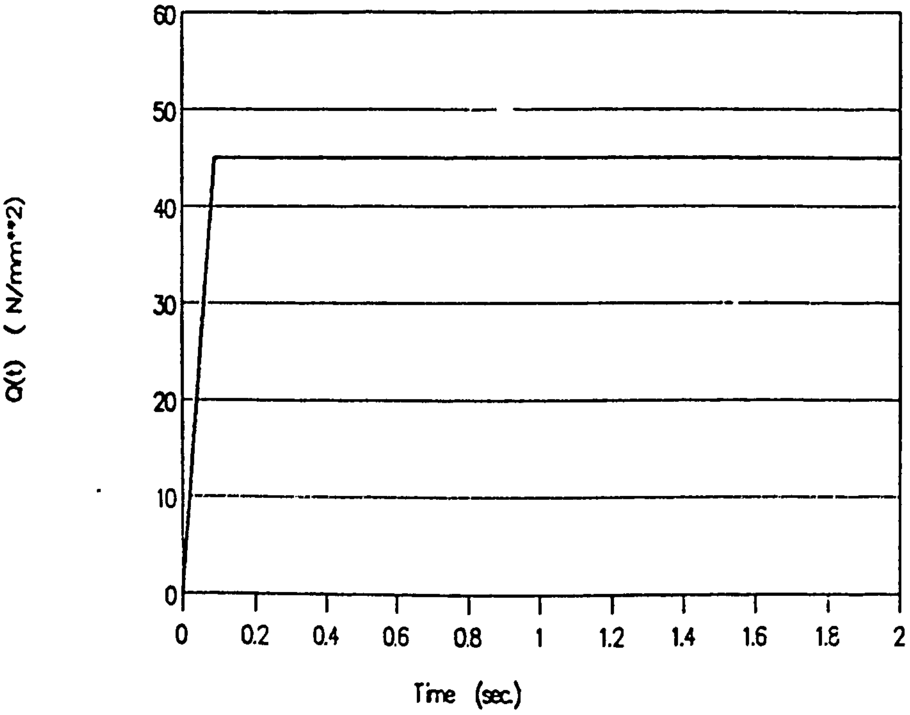
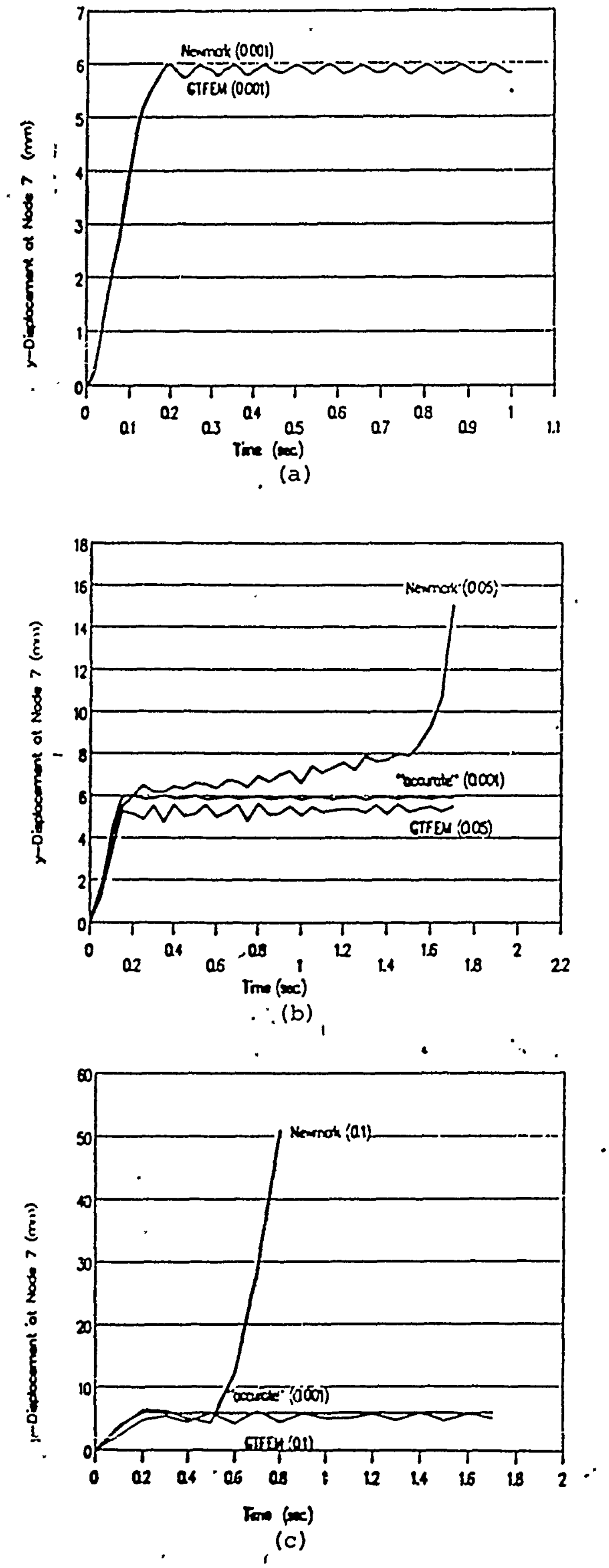


Fig. 19. Applied load for problem in Fig. 18



(a)

(b)

(c)

Fig. 20 Predicted Vertical Displacement vs. Time Plots for Newmark and GTFEM Schemes (a) Δt = 0.001 sec, (b) Δt = 0.05 sec, (c) Δt = 0.1 sec.

Center for Computational Mechanics
Washington University, St. Louis, MO 63130 USA
Technical Report WU/CCM-90/1

THE USE OF A PRIORI ESTIMATES IN ENGINEERING COMPUTATIONS†

B. A. Szabó
A. P. and B. Y. Greensfelder Professor of Mechanics
Washington University, St. Louis, MO 63130 USA

ABSTRACT

The use of a priori information in finite element analysis is discussed with reference to (a) definition of mathematical models; (b) selection of the extension process for controlling the errors of discretization; (c) a posteriori estimation of error in energy norm, and (d) extraction of engineering data from finite element solutions. The discussion focuses on the displacement formulation. Examples are presented.

1. INTRODUCTION

Corresponding to any well formulated mathematical model is an exact solution, denoted by $\vec{u}_{EX}$, which depends on the choice of formulation and the data which characterize the geometry, material properties, loading, and constraints, but is independent of the discretization. Corresponding to a particular choice of discretization is a finite element solution, $\vec{u}_{FE}$, and a number N, called the number of degrees of freedom, which is the number of linear algebraic equations one has to solve in order to obtain $\vec{u}_{FE}$. If the problem is well formulated and the discretizations are properly selected then $\vec{u}_{FE} \rightarrow \vec{u}_{EX}$ as $N \rightarrow \infty$.

In the following discussion only the displacement formulation is considered. In the displacement formulation the difference between $\vec{u}_{EX}$ and $\vec{u}_{FE}$ is naturally measured in the energy norm. By definition, the energy norm of a displacement

† *Presented at the Workshop on Reliability in Finite Element Computations, Austin, Texas, October 26-30, 1989. To appear in Computer Methods in Applied Mechanics and Engineering in 1990.*

function $\vec{u}$ is the square root of the strain energy of $\vec{u}$, which is usually denoted by $\|\vec{u}\|_E$. In the displacement formulation $\vec{u}_{FE} \rightarrow \vec{u}_{EX}$ in the following sense:

$$\lim_{N\rightarrow\infty} \|\vec{u}_{EX} - \vec{u}_{FE}\|_E = 0. \tag{1}$$

The choice of a particular discretization involves the creation of a set of functions S on the solution domain Ω. S is characterized by the mesh Δ, the mapping functions $\mathbf{Q}$, that is the functions which map standard finite elements onto the elements of the mesh, and the function spaces defined on the standard elements, called standard spaces. Most commonly the standard spaces are polynomial spaces characterized by the polynomial degree p. Hence $S = S(\Omega, \Delta, \mathbf{Q}, \mathbf{p})$. By definition, $\mathbf{Q} = \{\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \ldots, \mathbf{Q}^{\mathrm{M}(\Delta)}\}$, $\mathbf{p} = \{p_1, p_2, \ldots, p_{M(\Delta)}\}$ where $M(\Delta)$ is the number of elements in the mesh. In this paper only the case of uniform p, that is $p_i = p$ $(i = 1, 2, \ldots, M(\Delta))$ is considered.

The subset of functions in S which satisfies the kinematic boundary conditions is denoted by $\tilde{S}$. The number of degrees of freedom N is the number of linearly independent functions in $\tilde{S}$. The finite element solution is that function from $\tilde{S}$ which minimizes the error in energy norm:

$$\|\vec{u}_{EX} - \vec{u}_{FE}\|_E = \min_{\vec{u}\in\tilde{S}} \|\vec{u}_{EX} - \vec{u}\|_E. \tag{2}$$

The error $\|\vec{u}_{EX} - \vec{u}_{FE}\|_E$ can be reduced in various ways: by mesh refinement; changing the mapping; changing the standard spaces; increasing the degree of the standard polynomial space, or any combination of these. For reasons of implementation most commonly the mesh is refined, or the degree of the standard polynomial space is increased, or mesh refinement is combined with increase of the polynomial degrees. These approaches are respectively called *h-extension*, *p-extension* and *hp-extension*. When aspects of implementation rather than reduction of error are discussed then the word *version* instead of *extension* is used. From the point of view of implementation there are substantial differences between the h-version on one hand and the p- and hp-versions on the other.

2. THE DEFINITION OF MATHEMATICAL MODELS.

Mathematical models are essentially transformations. They transform one set of data, the input data, into another set, the output data. Mathematical models cannot improve on the quality of the information which resides in the input data. They can, however, damage the input information so that the output information is useless and worse, misleading. The most important use of priori analysis in engineering computations is related to the proper definition of mathematical models.

2.1. Models based on the displacement formulation.

For models based on the displacement formulation it is necessary that the potential energy of the exact solution $\Pi(\vec{u}_{EX})$ be bounded from below and the energy norm measure of the exact solution $\|\vec{u}_{EX}\|_E$ be nonzero. In addition, the model has to be consistent with the goals of computation: If the goal is to compute certain functionals, such as, for example, stress maxima, then the functionals of interest computed from the exact solution have to be finite.

Surprisingly, the problem of proper model definition has received very little attention in the finite element literature. Often model definitions are based on intuitively plausible reasoning which violates one or more of the conditions listed in the preceding paragraph. The most common modeling errors are:

1. Use of concentrated forces in two- and three-dimensional elasticity and in plate/shell models based on the Reissner-Mindlin or higher theories. This violates the condition that the potential energy must be bounded from below.
2. Use of point constaints. Point constraints should be used only as rigid body constraints, that is the external loads must satisfy equilibrium. If this condition is not satisfied then, depending on other boundary conditions, either $\Pi(\vec{u}_{EX}) = -\infty$, or $\|\vec{u}_{EX}\|_E = 0$, or $\vec{u}_{EX}$ is not distinguishable in the energy space from the solution which would be obtained if the point constraint were not applied.
3. One or more functionals of interest computed from the exact solution are infinite. For example, the goal is to determine stress maxima in elasticity and the exact value of the maximum stress is infinity.

The finite element solutions and/or the functionals of interest computed from the finite element solutions based on such models are entirely discretization-dependent. The data computed from the finite element solutions can be of credible magnitude and therefore very misleading. Some specific examples are presented in [1,2].

2.2. Models based on other formulations.

Models based on other than the displacement formulation must satisfy conditions analogous those described in Section 2.1 and, in addition, they must be shown to satisfy the Babuška-Brezzi condition [3]. The Babuška-Brezzi condition is an essential a priori requirement, satisfaction of which guarantees that the formulation works well for all admissible input data. (The displacement formulation trivially satisfies this condition). If the Babuška-Brezzi condition is not satisfied then the formulation will fail for some admissible input data. The practice of using formulations which do not satisfy this condition, has led to models that worked well in some cases but not in others. For example, elements based on such formulations may work well for the simple problems typically used in benchmark studies but fail in some of the more complicated problems used in professional practice. For some formulations it is very difficult to establish whether or not the Babuška-Brezzi condition is satisfied.

3. CONTROL OF THE ERRORS OF DISCRETIZATION

Given that h-, p- and hp-extensions are alternative approaches to controlling the errors of discretization, the question naturally arises: When and why should one choose h-extension, p-extension or hp-extension? This question is now examined from the theoretical and practical points of view. The essential difference between the theoretical and practical points of view is that theoretical estimates are concerned with the asymptotic behavior of the error measured in precisely defined norms, usually the natural norm of the formulation, whereas in practical analyses one is concerned with approximating certain functionals, typically to within one to five percent relative error. The functionals of interest may or may not be directly related to the natural norm, however they can be related to the natural norm by means of extraction procedures discussed in Section 5. Satisfactory approximations often can be achieved well before the asymptotic range is entered. Experience is an important a priori information for designing the discretization.

3.1. Classification of problems.

It is useful to classify the exact solution $\vec{u}_{EX}$ in relation to the finite element mesh into three categories:

Category A: $\vec{u}_{EX}$ is analytic† on each finite element, including the boundaries of each finite element.

Category B: $\vec{u}_{EX}$ is analytic on each finite element, including the boundaries of each finite element, with the exception of some of the vertices (in three dimensions also along some of the edges). The points where $\vec{u}_{EX}$ is not analytic are called *singular points.*

For problems in category B the exact solution in the neighborhood of singular points can be typically written as the sum of some smooth function and a function in the following form:

$$\vec{u}_{EX} = \sum_{i=1}^{\infty} A_i r^{\lambda_i} \vec{\Phi}_i(\theta) \qquad r > r_0 \tag{3}$$

where r, θ are polar coordinates centered on the singular point; A_i are coefficients which depend on the loading; λ_i is a fractional number, greater than

† *A function is analytic in a point if it can be expanded into a Taylor series about that point.*

zero; $\vec{\Phi}_i(\theta)$ is a smooth or piecewise smooth function, and r_0 is the radius of convergenece. It is possible to determine λ_i and $\vec{\Phi}_i$ from the definition of the model. Details are discussed, for example, in [4].

We will say that a problem is *strongly in category B* if $\lambda \overset{\text{def}}{=} \min \lambda_i < 1$ $(i = 1, 2, \ldots)$ otherwise it is *weakly in category B.* This choice of subdivision of category B is made with reference to the fact that in computational mechanics one of the usual goals is to compute stress data which are related to the first derivatives of the displacement. For problems strongly in category B the maximum stress is infinity in the singular point, whereas for problems weakly in category B the stress is finite on the entire domain. If the goals were to compute, say, the second derivatives of the stress, then it would be logical to regard problems for which $\lambda < 3$ as being strongly in category B. Whether a problem is strongly or weakly in category B often depends on decisions concerning idealization. Given alternative choices of idealization, it is generally preferable to select the idealization so that λ is as large as possible [5].

Category C: The mesh cannot be constructed so that singular points are at vertices (or, in three dimensions, the singular points are along element edges) or the locations where abrupt changes occur in the derivatives of $\vec{u}_{EX}$, such as material interfaces, are at interelement boundaries. This is because the locations of singular points are solution-dependent.

We will say that a problem is *strongly in category C* if $\lambda < 1$ and the distribution of singular points lacks any a priori recognizable pattern. In this case the solution is said to be *uniformly unsmooth.* We will say that a problem is *weakly in category C* if $\lambda \geq 1$ or the singular points are distributed in some regular pattern.

Let us consider, for example the tension strip with a circular hole, shown in Fig. 1. Let us assume that the material is modelled by the elastic-strain hardening stress-strain law, shown in Fig. 1, and the material exhibits the Bauschinger effect. If the material is never loaded such that the yield point σ_{yield} is exceeded then the solution is very smooth and the problem is very weakly in category B. The reason that it is not in category A is that in points A and B the solution is of the form

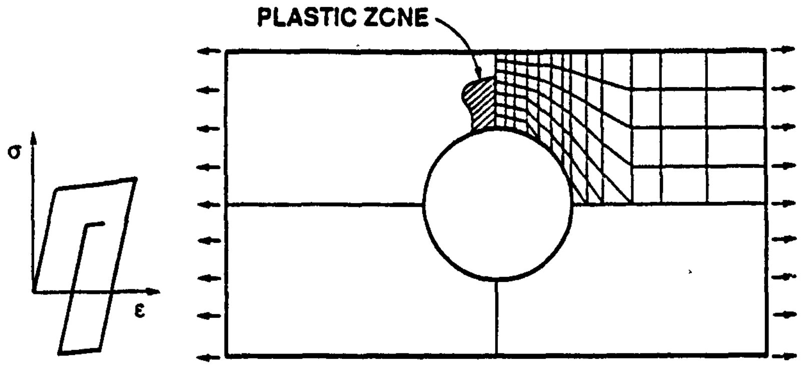


Fig. 1. Tension strip with a circular hole.

(3) with $\lambda \approx 2.75$, see for example [4]. If the load is increased past the yield point and kept constant, or cycled with a constant amplitude, the problem is weakly in category C. The singular points are the points on the boundary between the elastic and plastic regions which is, of course, solution-dependent. If the load is cycled with variable magnitude so that the yield stress is repeatedly exceeded in tension and compression then the problem is strongly in category C: the distribution of the singular points is dependent on the load history. In this case no pattern is discernible a priori for the distribution of singular points, except that there may be regions where the stress does not exceed the yield point, and the solution on those regions belongs in categories A or B.

3.2. A priori estimates of error.

Based on the classifications defined in Section 3.1, close a priori estimates are available for the error measured in energy norm. These are *asymptotic* estimates, that is the estimates are close when N is sufficiently large. A priori estimates indicate the rate of change of the error with respect to increasing number of degrees of freedom. Clearly, it would be better to know the rate of change of the error with respect to some work measure, such as the number of operations, rather than the number of degrees of freedom, but such work measures are strongly implementation- and machine-dependent, hence difficult to interpret.

Using the asymptotic rate of change of the error measured in energy norm with respect to the number of degrees of freedom as the basis for comparison, the available a priori estimates provide a basis for the choice of extension process:

For problems in category A the most effective method for controlling the errors of approximation is by p-extension because the error measured in energy norm, decreases *exponentially* when p-extension is used:

$$\|\vec{u}_{EX} - \vec{u}_{FE}\|_E \leq \frac{k}{\exp(\gamma N^{\theta})} \tag{4}$$

where k, γ, θ are positive numbers. If h-extension is used for problems in category A then the asymptotic rate of convergence is algebraic:

$$\|\vec{u}_{EX} - \vec{u}_{FE}\|_E \leq \frac{k}{N^{\beta}} \tag{5}$$

where k and β are positive numbers with $\beta = \min(p, \lambda)/2$.

For problems in category B the most effective method for controlling the errors of approximation is by hp-extensions: The mesh is graded so that the sizes of elements decrease toward the singular points in geometric progression with a common factor of about 0.15. Such meshes are called *geometric meshes.* The polynomial degree of elements is distributed linearly, with rounding to the nearest integer, such that the lowest polynomial degree is assigned to the smallest elements, the highest polynomial degree to the largest. If h- or p-extensions are used for problems in category B then the asymptotic rate of convergence is algebraic. If the singular points are nodal points then the asymptotic rate of convergence of p-extensions is characterized by $\beta = \lambda$, otherwise $\beta = \lambda/2$ in eq. (5). If the meshes are adaptively constructed for h-extensions then the asymptotic rate of convergence of h-extensions is characterized by $\beta = p/2$ otherwise by $\beta = \min(p, \lambda)/2$.

For problems strongly in category C h-extension is the best approach. In this case convergence is *algebraic.* For problems weakly in category C some combination of the h- and p-extensions or h- and hp-extensions is the best approach. In this case mesh refinement is used in those regions which contain the singular point and p- or hp-extension is used elsewhere. The asymptotic rate of convergence is algebraic but much faster rate of convergence can be realized in the preasymptotic range [6].

3.3. Practical aspects.

Most problems in linear elastostatics and linear elastodynamics and many nonlinear problems belong in category B and engineering accuracy can be achieved by p-extensions in the preasymptotic range if properly designed meshes are used. Hence the use of p-extensions and properly refined meshes are of substantial importance in engineering design and analysis. In fact, h-extension is not a good choice for such problems.

Particular choices of mesh depend on (1) what data are of interest; (2) what accuracy is desired, and (3) how the data are to be computed. For many problems in category B geometric refinement at singular points is not necessary and in many important cases even the requirement that a singular point has to be a mesh point can be relaxed: Engineering accuracy can be achieved with p-extension, and coarse finite element meshes. This is demonstrated by examples in Sections 6.

Whether sufficient accuracy can be achieved by p-extension for a particular choice of mesh cannot be known a priori. A priori information about the solution and the asymptotic rates of convergence can be used, however, in guiding the refinement process. In general, it is best to start with a simple mesh and perform p-extension. The error in energy norm is estimated by the method outlined in Section 4 and convergence of the quantities of interest is observed. If convergence of the quantities of interest is realized, and the error in energy norm is small, then the discretization error for the mathematical model is small. If convergence is not realized, and/or the error in energy norm is large, then the mesh has to be refined. For best results mesh refinement should be patterned after the optimal meshing for hp-extensions: The mesh should be graded in geometric progression toward the singular points with a common factor of about 0.15. p-Extension is then performed on this fixed geometric mesh. The results are checked for accuracy and, if necessary, additional layers of elements, graded in geometric progression, are introduced at the singular points. The number of layers needed thepend on the coefficient of the singular term. In most cases two layers are sufficient. Once again, a sequence of solutions is obtained by p-extension. The strain energy of the error decreases exponentially, provided that there is a sufficient number of elements at the singular points.

4. A POSTERIORI USE OF A PRIORI ESTIMATES.

The a priori estimates described in Section 2.2 can be used to obtain a posteriori estimates of error in energy norm if finite element solutions corresponding to a sequence of spaces $S_1 \subset S_2 \subset S_3 \ldots$ is available. Such sequences are naturally created by p-extensions and can be created by h-extensions also. The a posteriori use of a priori estimates are based on the following relationships: It can be shown that

$$\| u_{EX} - u_{FE} \|^2_{E(\Omega)} = \Pi_p - \Pi \tag{6}$$

where Π_p is the potential energy computed for polynomial degree p and $\Pi \overset{\text{def}}{=} \Pi(\vec{u}_{EX})$ is the potential energy corresponding to the exact solution [7]. If p-extension is used for problems in category B then the asymptotic rate of convergence is algebraic, that is

$$\Pi_p - \Pi \leq \frac{k^2}{N_p^{2\beta}}. \tag{7}$$

For sufficiently large N the 'less than or equal' sign can be replaced with 'approximately equal'. This is because (7) is a tight estimate for sufficiently large N. Using (7) for three finite element solutions corresponding to the spaces S_{p-2}, S_{p-1}, S_p, the constants k and β can be eliminated and the following relationship obtained:

$$\frac{\Pi - \Pi_p}{\Pi - \Pi_{p-1}} \approx \left(\frac{\Pi - \Pi_{p-1}}{\Pi - \Pi_{p-2}} \right)^Q \tag{8}$$

where Q depends only on N_{p-2}, N_{p-1}, N_p:

$$Q \overset{\text{def}}{=} \frac{\log(N_{p-1}/N_p)}{\log(N_{p-2}/N_{p-1})}. \tag{9}$$

To obtain an estimate of the exact potential energy Π, the nonlinear equation (8) has to be solved.

Although this estimator is asymptotically correct only when p-extension is used for problems in category B, computational experience has shown it to be reliable and generally accurate in the preasymptotic range and for problems in category A as well. For problems in category A the estimator tends to overestimate the error by a small margin. For problems in category B the estimator slightly overestimates the error in the preasymptotic range. It is highly accurate in the asymptotic range, but underestimates the error, by a small margin, in the the region of transition from the preasymptotic to the asymptotic range [7,8].

5. COMPUTATION OF ENGINEERING DATA FROM FINITE ELEMENT SOLUTIONS.

Having obtained a finite element solution, the problem is to determine the functionals of interest $\Phi_i(\vec{u}_{FE})$, such as displacements, stresses, stress intensity factors, natural frequencies, etc. so that these functionals are close to their exact values:

$$|\Phi_i(\vec{u}_{EX}) - \Phi_i(\vec{u}_{FE})| \le \tau_i |\Phi_i(\vec{u}_{EX})| \quad i = 1, 2, \ldots \tag{10}$$

where τ_i is some specified tolerance. In engineering computation τ_i is typically between one and five percent. There are various methods for computing $\Phi_i(\vec{u}_{FE})$ and there are large differences in how fast $\Phi_i(\vec{u}_{FE})$ approaches $\Phi_i(\vec{u}_{EX})$ as the number of degrees of freedom N is increased.

For example, stresses can be computed directly by first computing the strains from $\vec{u}_{FE}$, which involves computation of the derivatives of $\vec{u}_{FE}$, then using the stress-strain law to obtain the stresses. In the case of uniform or nearly uniform finite element meshes, stresses computed in Gauss points converge faster than stresses computed in other points. Such accelerated convergence is often called *superconvergence.* In many h-version finite element computer programs stresses are computed only in Gauss points and then smoothly extrapolated to the boundaries to take advantage of superconvergence. Extrapolation tends to underestimate the maximum stress, however, which typically occurs at the boundary. In other cases a priori information about the solution is used by incorporating the functional form of the solution in the finite element space S [9]. Accelerated convergence can be achieved also by means of special mapping procedures, such as quarter point mapping, or by problem-specific auxiliary mapping techniques [10].

It is also possible to compute $\Phi_i(\vec{u}_{FE})$ form the virtual work expression. This involves selection of a virtual displacement function $\vec{w}_i$ such that $\Phi_i(\vec{u}_{FE})$ is the virtual work corresponding to w. The specially chosen virtual displacement functions are called *extraction functions.* Specifically, denote the virtual work of internal stresses corresponding to the displacement field $\vec{u}$, due to the virtual displacement $\vec{v}$, by $B(\vec{u}, \vec{v})$ and the virtual work of external forces by $F(\vec{v})$. The finite element solution, $\vec{u}_{FE}$, corresponding to the principle of virtual work, lies in $\tilde{S}$ and satisfies:

$$B(\vec{u}_{FE}, \vec{v}) = F(\vec{v}) \quad \text{for all } \vec{v} \in S^{(0)} \tag{11}$$

where $S^{(0)}$ is the subset of functions in S which vanish on those boundary segments where kinematic boundary conditions are prescribed. Suppose that we can write:

$$\Phi_i(\vec{u}_{EX}) = B(\vec{u}_{EX}, \vec{w}_i) + Q(\vec{w}_i) \qquad (12a)$$

where $\vec{w}_i$ is an extraction function for Φ_i, $Q(\vec{w}_i)$ is a functional which is independent of $\vec{u}_{EX}$. Note the similarity with Green's function: If $\vec{w}_i$ is Green's function then $B(\vec{u}_{EX}, \vec{w}_i) = 0$ and the functional can be computed directly from the input data. The finite element approximation to $\Phi_i(\vec{u}_{EX})$ is:

$$\Phi_i(\vec{u}_{FE}) = B(\vec{u}_{FE}, \vec{w}_i) + Q(\vec{w}_i). \qquad (12b)$$

Therefore, from (12a,b),

$$\Phi_i(\vec{u}_{EX}) - \Phi_i(\vec{u}_{FE}) = B(\vec{u}_{EX} - \vec{u}_{FE}, \vec{w}_i)\,. \qquad (13)$$

The following a priori estimate is available for the error in Φ_i:

$$|\Phi_i(\vec{u}_{EX}) - \Phi_i(\vec{u}_{FE})| \leq \frac{1}{2}\|\vec{u}_{EX} - \vec{u}_{FE}\|_E \|\vec{z}_{EX} - \vec{z}_{FE}\|_E \qquad (14)$$

where $\vec{z}_{EX}$ is the solution of an auxiliary problem, which depends on the choice of $\vec{w}_i$ and $\vec{z}_{FE} \in S$ is the finite element approximation to $\vec{z}_{EX}$.. Details are available in [11,12,13]. In many cases it is possible to select $\vec{w}_i$ so that the error $\|\vec{z}_{EX} - \vec{z}_{FE}\|_E$ is not greater than $\|\vec{u}_{EX} - \vec{u}_{FE}\|_E$ and $\Phi_i(\vec{u}_{FE})$ converges to $\Phi_i(\vec{u}_{EX})$ at the same rate, or faster, than $\|\vec{u}_{EX} - \vec{u}_{FE}\|_E^2$. Thus superconvergence is realized. Specific case studies have been reported in [14,15,16,17,18].

The gains in efficiency realized by properly chosen extraction procedures are particularly important for three-dimensional problems where meshing for high accuracy is generally difficult and expensive. Superconvergent extraction procedures may be the only means by which reliable and accurate computation of engineering data can be made practical.

The a priori estimate of error (15) suggests that adaptive procedures designed to control the error in energy norm are directly applicable when computing functionals using the extraction procedure. Note, however, that adaptivity for multiple solutions is involved: Both $\|\vec{u}_{EX} - \vec{u}_{FE}\|_{E(\Omega)}$ and $\|\vec{z}_{EX} - \vec{z}_{FE}\|_{E(\Omega)}$ have to be small.

Depending on the goals of computation, there can be many extraction functions and therefore auxiliary problems.

Extraction methods can be implemented in both h- and p-version finite element codes, however the higher rate of convergence of p-extensions and the greater robustness of the p-version are inherited by the extraction methods and therefore the p-version codes will perform better.

6. EXAMPLE: ATTACHMENT LUG.

An attachment lug, typical of problems weakly in category B, is shown in Fig. 2a and the corresponding mesh in Fig. 2b. The lug is of uniform thickness of 0.5 in. Plane stress conditions are assumed. The modulus of elasticity is 3.0×10^4 k/in^2 and Poisson's ratio is 0.3. The goal is to find the magnitude and location of the largest tensile stress in the neighborhood of the circular hole when a sinusoidal normal pressure is applied on the inside perimeter of the hole, as shown in Fig. 2a. The pressure distribution $T_n = T_n(\theta)$ (k/in^2 units) is given by the expression:

$$T_n = \begin{cases} -\dfrac{32}{\pi}\cos\theta & -\dfrac{\pi}{2} \leq \theta \leq \dfrac{\pi}{2} \\ 0 & \text{otherwise.} \end{cases} \qquad (15)$$

The pressure distribution may vary such that the direction of its resultant (F = 10 k), characterized by the angle α, is in the range of ±45.0 degrees.

Generally a lug is part of a larger structural unit, such as a bulkhead in an airframe. It would be impractical to model the entire structural unit just to find the stress distribution in the neighborhood of the circular hole of the lug. For this reason some boundary condition is imposed on the lug along the interface between the lug and the structure to which it is attached. The choice of this boundary condition is a modelling decision which, together with the other input data, determines $\vec{u}_{EX}$. The underlying assumption is that the data of interest are not sensitive to the modelling decision. It is necessary to check whether this assumption is valid: A mathematical model cannot be considered as reliable if the data of interest are sensitive to arbitrary modelling decisions.

Two modelling decisions are compared in the following: First it is assumed that the lug is held in equilibrium by smoothly varying tractions imposed along the boundary AB. In this case the exact solution is weakly in category B. Second, it is assumed that the lug is fixed along AB. In this case the exact solution is strongly in category B. In reality, the lug is elastically constrained so that smaller deformation is allowed along AB than in the first case but larger than in the second.

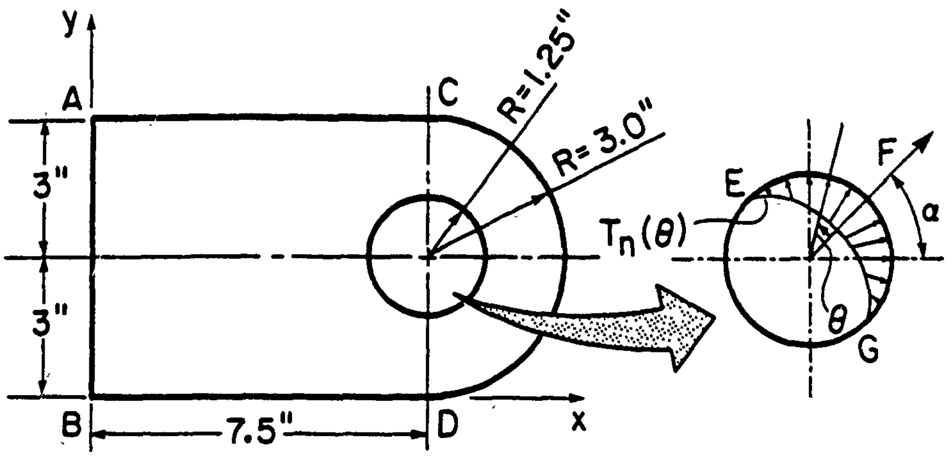


Fig. 2a. Attachment lug. Typical for problems weakly in category B.

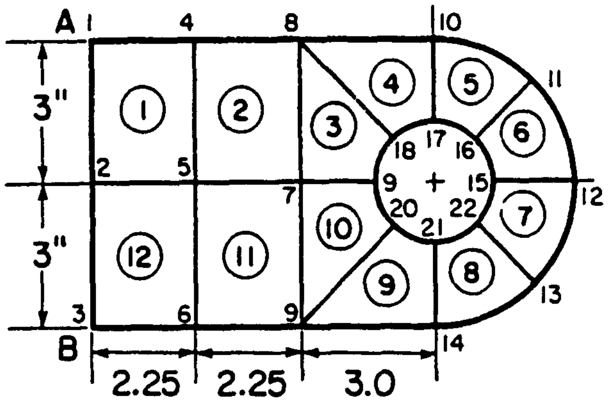


Fig. 2b. Finite element mesh for the attachment lug.

6.1. Equilibrium loading.

The lug is held in equilibrium by linearly distributed normal tractions and quadratically distributed shear tractions applied along AB:

$$T_x = -\frac{10}{3}\cos\alpha + \frac{25}{3}\sin\alpha(y-3) \tag{16a}$$

$$T_y = -\frac{5}{9}y(6-y)\sin\alpha. \tag{16b}$$

These tractions are consistent with the simple stress distributions of the usual engineering theory of beams. Only rigid body constraints are imposed.

The problem has several weakly singular points. These are points A, B, C, D, E, G, shown in Fig. 2a. The location of points E and G depends on the choice of α. To realize the faster asymptotic rates of convergence of p- and hp-extensions, singular points should be nodal points. It would be inconvenient, even impractical, to have a different mesh for each α, however. It is now demonstrated through a computational experiment that if the goal of computation is to determine the location and magnitude of the largest tensile stress to within one percent relative error then points E and G do not have to be nodal points. This loading has substantial practical importance because it is typical for loadings of pinned structural connections.

First, p-extension is performed on the mesh shown in Fig. 2b and the relative error in energy norm is estimated using the procedure described in Section 4. The relevance of this estimate for this particular problem is that the error in energy norm is closely related to the root-mean-square error in stresses, hence it is indicative of the average error in stresses [7]. The computations were performed with the computer program MSC/PROBE. The results, shown in Table 1, indicate that the relative error in energy norm is under one percent at $p = 8$.

Table 1. Attachment lug. Equilibrium loading, $\alpha = \pi/6$.
Estimated relative error in energy norm.

p	N	$\|\vec{u}_{FE}\|_E^2$ (in k)	Rel. Error (percent)
1	41	1.46368×10^{-2}	50.29
2	109	1.89925×10^{-2}	17.47
3	177	1.94095×10^{-2}	9.61
4	269	1.95218×10^{-2}	5.92
5	385	1.95680×10^{-2}	3.39
6	525	1.95857×10^{-2}	1.57
7	689	1.95891×10^{-2}	0.87
8	877	1.95900×10^{-2}	0.51

Second, the location and magnitude of the largest principal stress is computed and p-convergence is checked. These computations involve the use of a data mesh, constructed as follows: Each element in the mesh is related to the standard quadrilateral element by the mapping:

$$x = Q_x^{(k)}(\xi, \eta) \qquad y = Q_y^{(k)}(\xi, \eta) \qquad -1 \le \xi, \eta \le +1 \tag{17}$$

where k is the element number. Each of the intervals $-1 \le \xi \le +1$, $-1 \le \eta \le +1$ are subdivided into n subintervals, thereby creating an $n \times n$ data mesh on the standard quadrilateral element. The stresses are then computed in each nodal point of this data mesh for each element and the location and magnitude of the largest principal stress is identified with respect to the data mesh. Of course, the resolution depends on n and the size of the elements. For this example $n = 10$ was selected.

Table 2. Attachment lug. Equilibrium loading, $\alpha = \pi/6$.

Location and magnitude of the largest tensile stress (σ_1)

p	Element number	x (in)	y (in)	σ_1 (k/in^2)	σ_2 (k/in^2)	angle (degrees)
1	4	6.616	3.884	19.64	3.87	38.1
2	8	8.067	1.886	19.76	2.83	27.7
3	8	8.067	1.883	22.28	3.59	27.9
4	8	8.067	1.886	22.81	3.51	27.2
5	8	8.153	1.934	22.29	0.95	31.1
6	8	8.153	1.934	22.34	0.33	31.1
7	8	8.067	1.886	22.32	0.33	27.1
8	8	8.153	1.934	22.32	−0.21	31.3

The results, shown in Table 2, indicate that the location of the largest principal stress converges to a point which lies in the vicinity of points (8.067, 1.886) and (8.153, 1.934). Both points lie on the perimeter of the circular hole. Since the data mesh is characterized by $n = 10$, the angular resolution along the perimeter of the circular hole is 4.5 degrees (0.098 in). These points are very close to $\theta = -\pi/2$ from the line of action of the applied force which is shown in Fig. 2a. It is seen that the location and magnitude of the maximum principal stress are virtually independent of the discretization: They change very little with respect to incre sing p. The stress σ_2 is the minor principal stress. The angle between the direction of the largest principal stress and the x-axis is in the last column in Table 2. From these results it is possible to conclude that the largest principal stress occurs at the perimeter of the circular hole, its angular position is approximately $\theta = -\pi/2$ from the line of action of the applied force and its magnitude is approximately 22.3 k/in^2. Convergence is very fast: Good engineering accuracy is realized at $p = 3$. However, to verify that convergence has in fact occurred, extension must be continued for at least two p-levels beyond the p-level at which the desired accuracy has been reached. In this case p-extension could have been stopped at p=5.

Table 3. Attachment lug. Fixed along AB, $\alpha = \pi/6$.
Estimated relative error in energy norm.

p	N	$\|\vec{u}_{FE}\|_E^2$ (in k)	Rel. Error (percent)
1	38	1.44166×10^{-2}	50.64
2	102	1.86780×10^{-2}	19.15
3	166	1.91132×10^{-2}	11.93
4	254	1.92667×10^{-2}	7.94
5	366	1.93373×10^{-2}	5.16
6	502	1.93673×10^{-2}	3.35
7	662	1.93773×10^{-2}	2.47
8	846	1.93822×10^{-2}	1.88

On comparing the results in Tables 1 and 2 it is seen that the error in the maximum tensile stress is much smaller than the error in energy norm. The error in energy norm is related to the root-mean-square error in stresses over the entire domain. Locally the stress can be less accurate or, as in this case, more accurate.

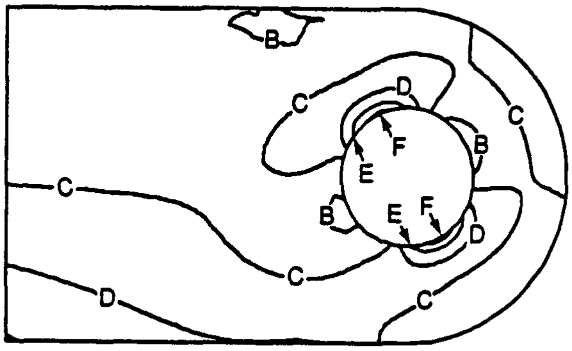


Fig. 3. Attachment lug. Contours of the first principal stress, $\alpha = \pi/6$.
Contour interval: 5.0 k/in^2, $B = 0$, $F = 20.0$ k/in^2.

A contour plot of the largest principal stress, generated from stress values computed directly form the finite element solution in the nodes of a 10×10 data mesh per element, is shown in Fig. 3.

6.2. Zero displacement along AB.

In this case the problem is strongly in category B. The computed values of the strain energy and the estimated relative error in energy norm are shown in Table 3. It is seen that convergence is somewhat slower, nevertheless reasonably good accuracy is achieved in energy norm.

Search for the location and magnitude of the maximum principal stress, restricted to elements 3 to 10, indicated strong convergence and yielded virtually the same results as those listed in Table 2. Hence the computed data are insensitive to both the modelling assumptions and the discretization.

6.3. Cracked attachment lug.

Let us now assume that a crack, 0.5 inches long, has developed in the lug, as shown in Fig. 4. The same kind of sinusoidal loading is applied as before, however in this case $\alpha = \pi/4$. The goal of computation is to determine the mode I and mode II stress intensity factors, respectively denoted by K_I, K_{II}. Two extraction methods, called cutoff function method (CFM) and the contour integral method (CIM), were used, detailed description of which is given in [14]. The mesh is now modified so that it is geometrically graded at the crack tip, which is typical for hp-extensions, however p-extension is used on this mesh. This mesh is comprised of forty elements.

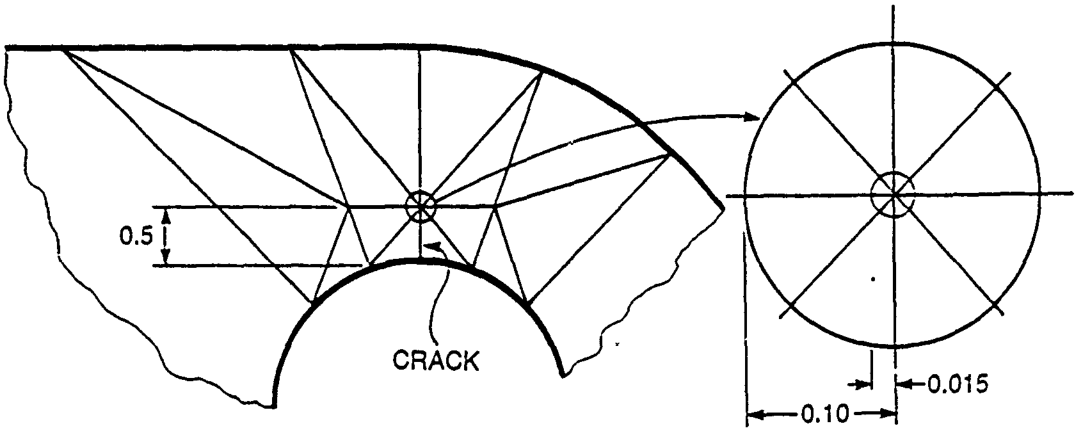


Fig. 4. Cracked attachment lug. Mesh detail.

The results given in Table 4 show that the stress intensity factors computed by the cutoff function method (CIM) and the contour integral method (CFM) methods converge strongly and obviously, although not monotonically. Greater accuracy and more nearly monotonic convergence is exhibited by the cutoff function method than the contour integral method. Both methods yield solutions which are within the range of precision normally needed in engineering computations at $p = 4$. The data for $p > 4$ merely confirm that convergence has in fact occurred.

Table 4. Cracked attachment lug, 40 finite elements.

p-Convergence of stress intensity factors in ksi$\sqrt{\text{in}}$ units.

p	N	$K_I/\sqrt{2\pi}$ (CIM)	$K_I/\sqrt{2\pi}$ (CFM)	$K_{II}/\sqrt{2\pi}$ (CIM)	$K_{II}/\sqrt{2\pi}$ (CFM)
1	93	5.335	4.252	−2.010	−1.607
2	269	3.492	3.851	−1.715	−1.866
3	473	3.954	3.836	−1.924	−1.842
4	757	3.731	3.785	−1.818	−1.845
5	1121	3.802	3.779	−1.852	−1.840
6	1565	3.773	3.783	−1.836	−1.841
7	2089	3.789	3.785	−1.845	−1.842
8	2693	3.783	3.785	−1.842	−1.843

7. NOTES ON IMPLEMENTATION.

From the point of view of implementation there are several important differences between the h- and p-versions. One of these is the difference in mapping requirements: In the p-version the size of the elements is not reduced as the degrees of freedom are increased, hence the geometric description must be independent of the number of elements. For this reason mapping by the blending function method is used. In the case of the lug problem the circular boundaries were represented exactly in the computation of the stiffness matrices and load vectors.

For the same reason, representation of the loading must be independent of the number of elements. In the example of the attachment lug, the loading was specified by the formulae (15) and (16a,b). Terms of the load vector were computed by Gaussian quadrature, using 12 Gauss points for each loaded element side. When an element is only partially loaded, as in the case of $\alpha = \pi/6$ elements 4 and 8 are, then the normal pressure is zero in some of the Gauss points. This causes some perturbation in the integrand. Nevertheless, as seen in this example, this affected neither the overall quality of the solution, measured in energy norm, nor the accuracy of the largest principal stress significantly. The formula is most conveniently defined in a local coordinate system, rotated by the angle α in relation to the global system. In this way the orientation of the applied load can be changed by changing only a single input parameter α.

In the h-versions stresses are usually evaluated at Gauss points only. For other points the stresses are determined by some smooth interpolation. There is no advantage in doing this in the p-version. In this example, the stresses were evaluated on the data mesh directly. No smoothing or averaging was performed. In fact, at interelement boundaries the stresses were evaluated independently for each element. The smoothness of the contour lines in Fig. 3 indicates that there are no significant jumps in the computed stress values between adjacent elements, which is another indicator that the solution is of good quality.

8. SUMMARY.

Mathematical models are reliable if the error measured in the natural norm of the formulation is small and the data of interest [illegible]e insensitive to both the choice of discretization and the modelling assumptio[illegible]ance of reliability is a systematic process in which sensitivities to discret[illegible]ion and modelling assumptions are investigated. In practical engineering decision-making processes the elapsed time between a problem being stated and some decision having to be rendered is generally quite limited, hence investigation of sensitivities is feasible only if the model is efficient. For this reason reliability and efficiency are closely related in practical computations.

The use of a priori information about the solution plays a very important role in assuring the reliability of mathematical models. In connection with displacement formulations four areas of application of a priori information were discussed in this paper:

1. Proper definition of mathematical models. It was noted that intuitively plausible reasoning can lead to conceptually flawed models and misleading results. Some basic knowledge on the part of the analyst about what data are admissible for a given formulation and goals of computation is essential.
2. Proper selection of the discretization is based on a priori classification of the solution. The important differences between the theoretical (asymptotic) estimates of performance and practical (preasymptotic) performance was emphasized. In p-extensions the desired accuracy should be achieved in the preasymptotic range. This is possible by proper a priori mesh design combined with a posteriori error estimation and, if necessary, modification of the mesh.
3. Error estimation. Although a priori estimates give information only about the asymptotic rate of convergence, they can be used a posteriori to estimate the relative error in energy norm.
4. Extraction. Engineering data can be computed from finite element solutions very efficiently through the virtual work expression. This requires proper selection of the extraction function which is based on a priori information about the solution.

ACKNOWLEDGEMENT.

Support of the author's work by NASA Langley Research Center under Grant No. NAG1-639 is gratefully acknowledged.

REFERENCES.

[1] Szabó, B. "On Errors of Idealization in Finite Element Analyses of Structural Connections", Proceedings, Workshop on Adaptive Methods for Partial Differential Equations, edited by J. E. Flaherty, P. J. Paslow, M. S. Shephard and J. D. Vasilakis, Society for Industrial and Applied Mathematics, Philadelphia pp. 15-28 (1989)

[2] Bortman, J. and Szabó, B. A., "Structural Analysis of Fastened Joints", Proc., 1989 US Air Force Structural Integrity Program Conference, San Antonio, TX, Dec. 5-7, (1989).

[3] Brezzi, F., "A Survey of Mixed Finite Element Methods", *Finite Elements. Theory and Applications*, edited by D. L. Dwoyer, M. Y. Hussaini and R. G. Voigt, Springer Verlag, New York, pp. 34-49 (1988).

[4] Williams, M. L., "Stress Singularities Resulting from Various Boundary Conditions in Angular Corners of Plates in Extension" *Journal of Applied Mechanics*, Vol. 19, pp. 526-528 (1952).

[5] Szabó, B. A., "Geometric Idealizations in Finite Element Computations", *Communications in Applied Numerical Methods*, Vol. 4, pp. 393-400 (1988).

[6] Oden, J. T., "Progress in Adaptive Methods in Computational Fluid Dynamics", *Adaptive Methods for Partial Differential Equations"*, J. E. Flaherty, M. S. Shephard and J. D. Vasilakis, editors, Society for Industrial and Applied Mathematics, Philadelphia pp. 206-252 (1989).

[7] Szabó, B. A. and Babuška, I., *Finite Element Analysis*, Manuscript of book to be published by John Wiley in 1990.

[8] Szabó, B. A., "Mesh Design for the p-Version of the Finite Element Method", Computer Methods in Applied Mechanics and Engineering, Vol. 55, pp. 181-197 (1986).

[9] Atluri, S. N., "Higher-order, Special and Singular Finite Elements" *State of the Art Survey of Finite Element Technology*, edited by A. K. Noor and W. Pilkey, ASME, New York (1983).

[10] Babuška, I. and Oh, H.-S., "The p-Version of the Finite Element Method for Domains with Corners and for Infinite Domains" Technical Note BN-1091, Institute for Physical Science and Technology, University of Maryland, College Park, MD (1988).

[11] Babuška, I. and Miller, A., "The Post-Processing Approach in the Finite Element Method - Part 1: Calculation of Displacements, Stresses and Other Higher Derivatives of the Displacements", *Int. J. num. Meth. Engng.*, Vol. 20, pp. 1085-1109 (1984).

[12] Babuška, I. and Miller, A.,"The Post-Processing Approach in the Finite Element Method - Part 2: The Calculation of Stress Intensity Factors", *Int. J. num. Meth. Engng.*, Vol. 20, pp. 1111-1129 (1984).

[13] Babuška, I. and Miller, A., "The Post-Processing Approach in the Finite Element Method - Part 3: A-Posteriori Error Estimates and Adaptive Mesh Selection", Int. J. num. Meth. Engng., Vol. 20, pp. 2311-2324 (1984).

[14] Szabó, B. A., "Solution of Electrostatic Problems by the p-Version of the Finite Element Method", Communications in Applied Numerical Methods, Vol. 4, pp. 565-572 (1988)

[15] Szabó, B. A. and Babuška, I., "Computation of the Amplitude of Stress Singular Terms for Cracks and Reentrant Corners", *Fracture Mechanics: Nineteenth Symposium, ASTM STP 969*, T. A. Cruse, Ed., American Society for Testing and Materials, Philadelphia, pp. 101-124 (1987).

[16] Izadpanah, K. "Computation of Stress Components in the p-Version of the Finite Element Method" Doctoral Dissertation, Washington University in St. Loius, MO (1984).

[17] Vasilopoulos, D., "Treatment of Geometric Singularities with the p-Version of the Finite Element Method", Doctoral Dissertation, Washington University in St. Louis, (1984)

[18] Papadakis, P. "Computational Aspects of the Determination of the Stress Intensity Factors for Two-Dimensional Elasticity", Doctoral Dissertation, University of Maryland, College Park, MD (1988).

The Problem of Modeling the Elastomechanics in Engineering

Ivo Babuška
Institute for Physical Science and Technology
University of Maryland
College Park, MD 20742 USA

Partially supported by the Office of Naval Research under Grant N00014-89-J-1030, and by the National Science Foundation under Grant CCR-88-20279.

Abstract

The paper describes the major aspects of modeling engineering problems of elastomechanics. It shows various aspects and results on a set of illustrative examples of 2 and 3 dimensional problems.

1. Introduction

The aim of computational analysis is to describe and *reliably predict* physical phenomena of interest. In the engineering sciences the primary aim is usually to design tools which operate SAFELY under certain (mechanical) conditions, in certain environments and for a certain period of time.

By computational analysis, ONLY mathematical problems and NOT the reality can be analyzed. The mathematical problem TRANSFORMS given input data into information which is of direct interest and does not add anything new (in fact its loses some information).

The aim of computation is to reliably obtain certain information in the range of an admissible tolerance so that it is not unduly influenced by the computational procedure used.

The formulation of the mathematical problem is usually the most crucial part of the analysis. Because of the complexity of engineering analysis and uncertainties in the available information, the formulation of the mathematical problem is often directly or indirectly stipulated in the design codes and often (at least in parts) it is also influenced by the particular (company) engineering practices. These codes change with time and express the experiences with the technology used. As a typical example we mention the design code (USAF-MIL-A-83444) used in aircraft components. It is based on the principle of "non-inspectable slow crack growth" which should meet the following demands

a) the life of the component should exceed two design times

b) the residual strength of the component should, after being in service two design life time, exceed maximal load acting on the component by a factor of, say, 9/8.

These or similar principles and other considerations (for example, uncertainties in the input information) lead to the *precise formulation of the mathematical problem and defining the data which have to be obtained* as well as to the admissible accuracy with which they have to be determined.

The basic flow chart of an engineering computational analysis is shown in Fig. 1.

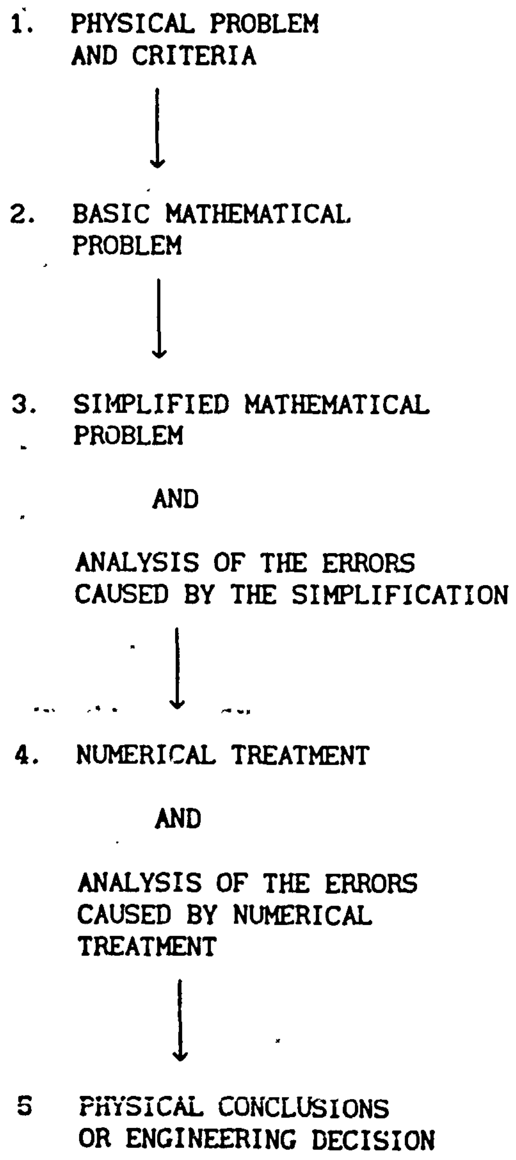

1. PHYSICAL PROBLEM AND CRITERIA
2. BASIC MATHEMATICAL PROBLEM
3. SIMPLIFIED MATHEMATICAL PROBLEM
AND
ANALYSIS OF THE ERRORS CAUSED BY THE SIMPLIFICATION
4. NUMERICAL TREATMENT
AND
ANALYSIS OF THE ERRORS CAUSED BY NUMERICAL TREATMENT
5 PHYSICAL CONCLUSIONS OR ENGINEERING DECISION

Fig. 1. The flow chart of computational analysis.

Usually in practice loops are present in the flow chart.

The "reality" is associated with (1). The engineering analysis of the problem, the aims of the analysis, and the assessement of the quality of the available data, etc. then yield the precisely formulated mathematical problem (model) (2). This model is to be understood as a "higher" model *which is identified with reality*. Nevertheless we solve usually only a simplified problem (3) and reliability of its solution is judged typically in comparison with (2). In (4) we solve numerically problem (3) and the reliability and the error of the numerical solution is related to the (exact) solution of (3) (and not (2) or (1)).

Let us underline that basic and simplified mathematical problem has to have reasonable mathematical properties, for example the existence of a solution. The existence of the solution of the "real" problem does not necessarily mean that the solution of the *mathematical problem* exists too. This is because of the simplification which enters into the formulation of the mathematical problem. Also, if the numerical algorithm provides numerical results (possibly reasonable looking), it does not mean that the solution of the mathematical problem necessarily exists, (because convergence has not to occur etc.) Obviously, theoretical analysis of the mathematical problem and comparison of its properties with the expected properties of the reality is essential part of the reliability of the model.

In all stages we have to relate the numerically obtained solution to the exact solution of a mathematical problem. The agreement with reality, for example with experiments, is then related solely to the formulation of the basic (or possibly simplified) problem. It is essential that the errors of numerical solutions are completely under control so that the exact solution of the mathematical problem is essentially achieved and a possible disagreement with experiments is related ONLY to the mathematical problem itself.

In this paper we will show various concrete examples to illustrate the basic ideas and results. All the computations in this paper has been made by the h-p version of the finite element method by the code PROBE (McNeil Schwendler-Noetic) and STRIPE (Aeronautical Institute of Sweden). These codes have various error checks so that the numerical results presented here can be assumed to be exact in the range of accuracy needed for the model conclusions.

2) Problem of the cantilever beam

Consider a problem of a simply supported cantilever beam shown in Fig.2.1a

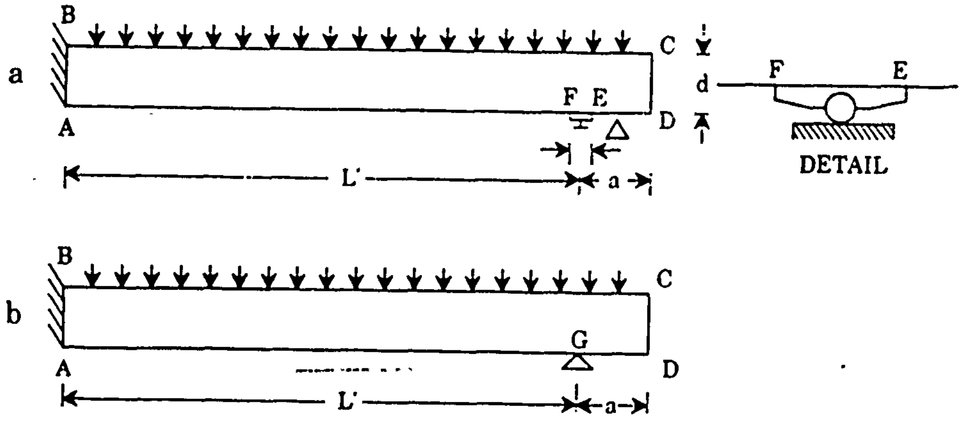


Fig.2.1 The simply supported cantilever beam

Let *the basic mathematical problem* be the problem of two dimensional linear elasticity (plane strain) for isotropic homogeneous material. The basic unknowns, the displacement u,v, satisfy the usual Lamé-Navier equations of elasticity. As the boundary conditions shown in the Fig.2.1a. we impose on $\overline{BC}: T_y = p, T_x = 0$, on $\overrightarrow{AB}$, $u = v = 0$, on $\overline{CD}$, $\overline{DE}$, $\overline{FA}: T_x = T_y = 0$ and on $\overline{EF}$(where $\Delta \ll d$) $T_x = 0$, $v = b(x-L')$ where b is such that $\int_{EF} T_y(x-L')dx = 0$.

By T_x respect to T_y we denoted the tractions. The problem is a model of the simply supported cantilever beam. The (weak) solution which has finite energy *exists and is uniquely determined.*

Let us now consider the *simplified problem* when $\Delta = 0$ and when $v = 0$ at the point G (see Fig.2.1b) is prescribed instead of the more complicated boundary condition of the basic problem. Then it is possible to show that the unique (weak) solution of the *simplified problem is the same* as the one when the condition $v = 0$ at G *is not present*. We see that the solution of the simplified problem is unacceptable. The reason for it is that the displacement under concentrated load is infinite (the Bousinesque solution). Although the solutions of the basic problem converges to the simplified one as $\Delta \longrightarrow 0$, the convergence *is very slow* and so it is inadmissible to consider this limiting case instead of the original one.

We remark that the point support is standardly used in finite element computations, i.e. the simplified problem is often numerically solved. Hence the error of the finite element solution is very large because for a mesh not extremely refined the cantilever beam solution without support, i.e. the exact solution is not obtained. It is possible to show that the finite element solution converges to the solution for the beam without support as the mesh size converges to zero. Hence the solution obtained in practice is mesh dependent. This is of course completely undesirable. It is necessary to mention that the finite element meshes used in practice (if not adaptively constructed) are crude and the FE solution does not show the mentioned effect. For some numerical analysis and computation we refer to [3].

3) The problem of the built in plate (beam).

Let us consider the classical problem of an infinite plate problem (in 2 dimension) which can be formulated as two dimensional (plane strain) problem in the coordinates xy. The scheme of a concrete example is shown in the Fig.3.1 where we assume built in (clamped) boundary conditions.

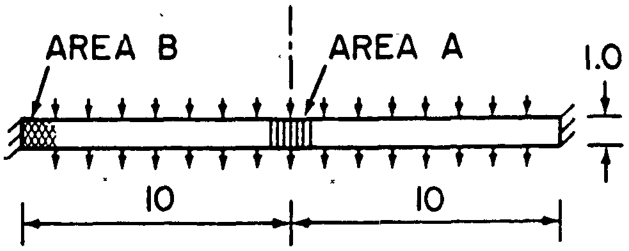


Fig.3.1 The scheme of the considered plate.

We define once more *the basic problem* as the two dimensional elasticity problem with $\nu = 0.3$ (where we denoted by ν the "Poisson ratio") and with the modulus of elasticity $E = 3.10^7$. The built-in (clamped) boundary condition is modeled by $u = v = 0$.

The solution exists, and is unique. Assume that the *aim* of the analysis *are the stresses* in the areas A and B shown in Fig. 3.1. Let us further distinguish 2 cases for the data of interest

a) the bending moment and the shear force

b) maximal stresses and the stress distribution through a cross section.

3.1. The problem of the boundary conditions.

It is obvious that the modeling of the clamped end *is an idealization*. In reality the support is obviously more complex. Hence we have an uncertainty in the formulation of the boundary condition and the problem shown in Fig.3.1 can be understood as the simplified one. To analyze this problem consider a few configurations which could be expected to lead the same simplified problem. They are shown in Fig.3.2.

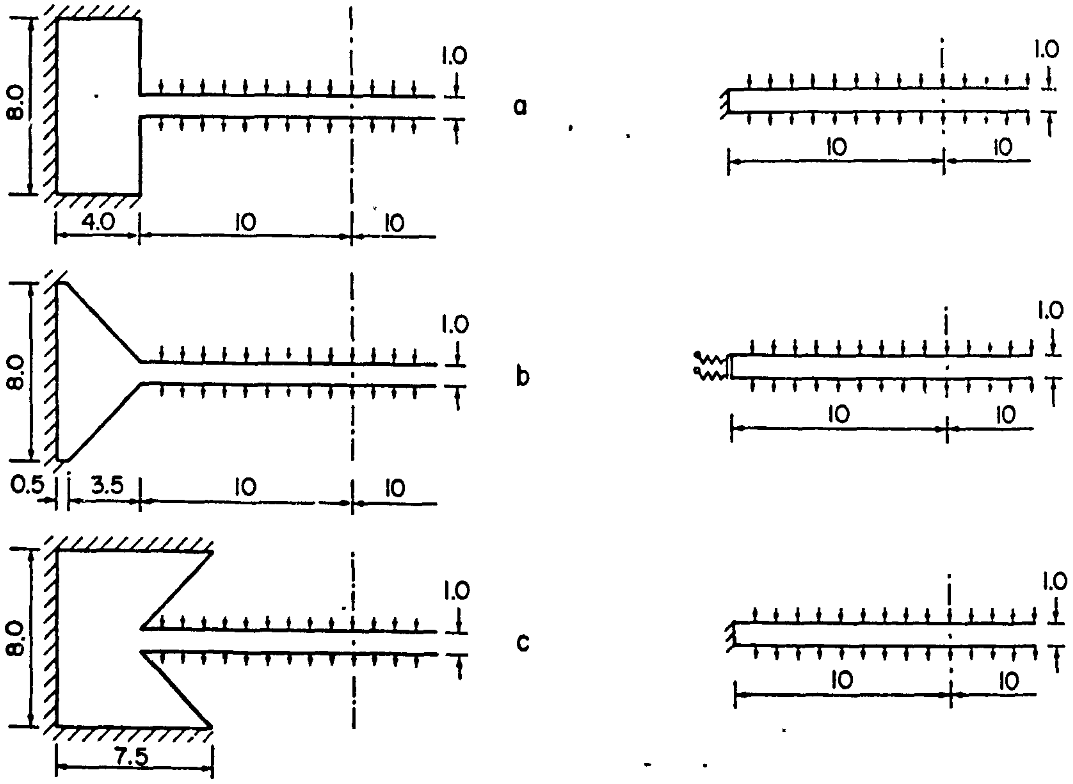


Fig.3.2 The scheme of various boundary conditions.

In the case (e) we model the clamped end as the elastically built-in end. The boundary conditions are then

$$T_x = -cv$$

$$T_y = -cw$$

where $c = 10^8$.

The case (f) depicts a still further simplified problem based on the Kirchhoff beam theory (strength of material approach). In this case the value of the bending moment in the center is

$$M = \frac{1}{24} 20^2 = \frac{100}{6} \tag{3.1}$$

Assume first that we are interested in the maximal stress σ_x *in the center of the beam* (area A in Fig.3.1). We get then in the case f : σ_x = 100.

We let the stress $\sigma_x = 100.23$ in the case d) (i.e. the case shown in Fig.3.1) be the "exact" solution to which we will compare all others. Table 3.1 shows the results.

Table 3.1. The stresses at the center for various models.

Case	σ_x	Error
a	108.46	8.2%
b	109.76	9.5%
c	108.27	8.0%
d	100.23	0%
e	120.17	19.9%
f	100.00	0.2%

Table 3.1 shows that simplification of either problem a,b,c leads to the error about 10% while the simplification of the case d by Kirchhoff hypotheses leads to error of 0.2%.

We have considered the special case for the ratio $d/L = \frac{1}{20}$. If $d \to 0$ (for fixed L), the relative difference between the cases a,b,c, and d goes to zero. We have then

In the case a):

$$(3.2) \qquad \alpha_{ad}(d) = \frac{|\sigma_x^a - \sigma_x^d|}{|\sigma_x^d|} = c_1 d + \text{higher order terms.}$$

For $d/L \le \frac{1}{20}$ we can neglect in (3.2) the higher order terms. Table 3.2

shows that in fact $\alpha_{ad}(d) = c_1 d$ with high accuracy.

Table 3.2 . The relative error $\alpha_{a,d}(d)$ of the case a with respect to the case d

d/L	$\alpha_{ad}(d)$
1/20	8.2%
1/50	3.2%

Let us now consider simplification of the problem d to the problem f. We have then

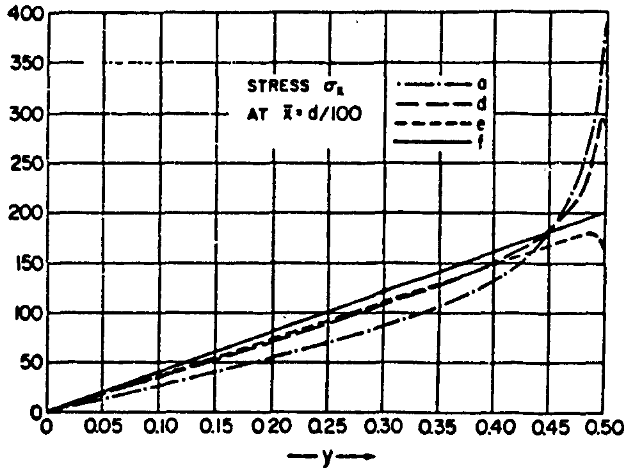
$$\beta_{df}(d) = \frac{|\sigma_x^d - \sigma_x^f|}{|\sigma_x^d|} = c_2 d^2 + \text{higher order terms.}$$

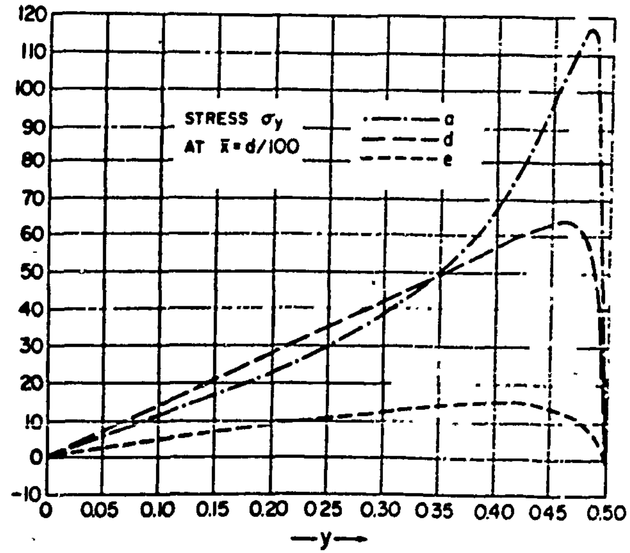
We see that the error of the modeling of the boundary conditions is much more significant than the error of the simplification leading to the Kirchhoff (strength of material) solution.

So far we have computed the maximum of the stress σ_x in the area A. Here the stress is very accurately linear through the cross section and hence when the interest is in the moments, the relative errors mentioned in the table 3.1 and 3.2 hold too.

We reported the stress in the area A and have seen that the sensitivity to the boundary conditions is of order 10%.

In the area B the differences in the stresses are much larger. The stresses are singular (the singular behavior of the solution will be discussed in the next section). Here we report in the Fig.3.3 a,b,c, the stress in the cross section in the distance $\bar{x} = d/100$ from the boundary. We clearly see that the differences between the mentioned cases are significant. On the other hand let us be interested in the (bending) moment and shear force. Then using equilibrium condition we easily see that the differences in the moments are the same as in the center of the beam i.e. as in the area A. Hence different aims lead to very different sensitivities to the uncertainties in input data.





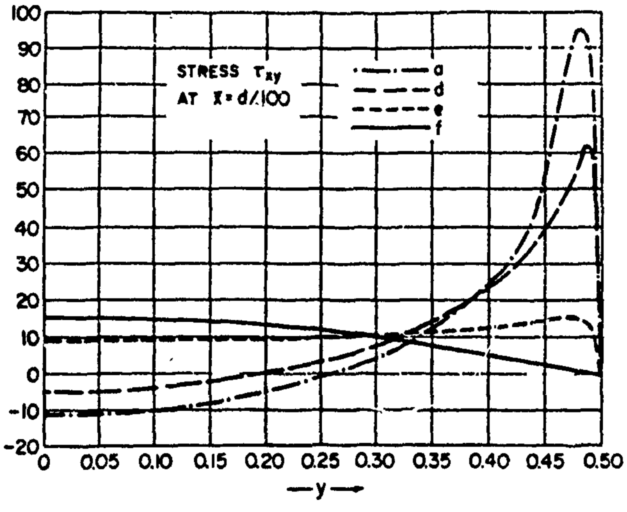


Fig. 3.3 The stresses in the area B.

4) The singularity problem and zooming principles.

4.1 The problem in 2 dimensions.

Let us consider the linear elasticity problem on a polygon domain Ω with the boundary consisting of straight segments Γ_i, $i = 1,\ldots,n$ and vertices $A_i, i = 1,\ldots,n$. By ω_i we denote the internal angles. Let us assume that we are dealing with a homogeneous isotropic material and that on every segment Γ_i a boundary conditions with analytic data prescribed and that no volume forces are present. Then the solution is analytic on $\bar{\Omega}\setminus\cup A_i$.

The solution is singular in the neighborhood of the vertices. Let Γ_i, Γ_{i+1} be the segments meeting in the vertex A_i. Assume for simplicity that the boundary condition (of standard type) are homogeneous on Γ_i, Γ_{i+1}.

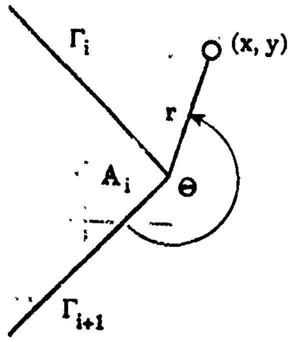


Fig. 4.1 Scheme of the angle

Then the solution of the problem in the neighborhood at A_i has the form

$$\text{(4.1)} \qquad \left.\begin{matrix} u(x,y) \\ v(x,y) \end{matrix}\right\} = \sum_{i=1}^{n} c_i r^{\lambda_i} \left\{\begin{matrix} \varphi_i(\theta) \\ \psi_i(\theta) \end{matrix}\right. + \text{ smoother terms}$$

where (r,θ) are the polar coordinates with the origin in A_i as shown in Fig. 4.1. λ_i are real a complex and $\operatorname{Re}\lambda_{i+1} \geq \operatorname{Re}\lambda_i > 0$ and φ_i and ψ_i are smooth functions in θ. If λ_i is complex then we use real and imaginary parts separately. Coefficients λ_i and φ_i, ψ_i are given by the geometry (the angle ω_i), the type of boundary condition on Γ_i, Γ_{i+1} and the material properties (in our case Poisson ratio). They are independent of the solution. The coefficients c_i depend *globally* on the solution (except for special cases). The structure of the solution is well known in the general case, for straight and curved segments Γ_i, and for general (linear) materials. We will not go into details. Here we refer instead to [14], [15], [16], [22]. In [21] a general approach (and a computer code) for computation of λ_i, φ_i, ψ_i for anisotropic and nonhomogeneous materials is given. Very often the coefficients c_i called stress intensity factors (together with λ_i, φ_i, ψ_i,) are the main aims of the analysis (in 2 and 3 dimensional settings). This is, for example, in the case of design based on the earlier mentioned design code

(USAF-MIL-A-83444). The stresses in the neighborhood of A_i are unbounded when $\mathrm{Re}\lambda_1 < 1$ provided that $c_1 \neq 0$.

Because in finite element computations the stresses are always finite, the character of the computed stresses can be misleading. If $\lambda_1 < 1$ then practically always (except in symmetric cases) $c_1 \neq 0$ although it can be *relatively* small. Then the large stress can be confined to small area only (we will see an illustration in a three dimensional problem in the next section.) *Reliable computational analysis always requires to compute these stress intensity factors.* (For methods of reliable computation of the singular behavior around the corners see references [1], [25]).

Let us now relate (4.1) to the *zooming principle.* To this end let $\xi = \frac{x}{\kappa}$, $\eta = \frac{y}{\kappa}$, $\rho = \frac{r}{\kappa}$ and

$$U(\xi,\eta) = u(\kappa\xi,\ \kappa\eta)$$

$$V(\xi,\eta) = v(\kappa\xi,\ \kappa\eta)$$

be the zoomed solution. Then we obviously have

$$\begin{matrix} U(\xi,\eta) \\ V(\xi,\eta) \end{matrix} = \left\{ \sum_{i=1}^{n} c_i F_i(\kappa) \quad \rho^{\lambda_i} \left\{ \begin{matrix} \varphi_i(\theta) \\ \psi_i(\theta) \end{matrix} \right. \right. \tag{4.2}$$

and hence functions $\rho^{\lambda_i}\varphi_i(\theta)$, $\rho^{\lambda_i}\psi_i(\theta)$ are the parts of the zoomed solution with the zooming parameter κ. Here it was characteristic that the infinite sector was invariant with respect to the zooming.

Let us consider now another case, namely the case shown in Fig. 4.2. We can zoom (Fig. 4.2b) the solution at A (see Fig. 4.2a) and get the solution in the form (4.1) (resp. (4.2)) in the same way as before. The other possibility is to zoom the solution with respect to the parameter d as shown in Fig. 4.2b. Then up to rigid body motion the first term in the zoomed solution has the form

$$(4.3)\qquad \begin{matrix} U(\xi,\eta) \\ V(\xi,\eta) \end{matrix} = d^{-1}M \begin{cases} \Phi_1(\xi,\eta) \\ \psi_1(\xi,\eta) \end{cases} + T \begin{cases} \Phi_2(\xi,\eta) \\ \psi_2(\xi,\eta) \end{cases} .$$

Here M and T is the moment and shear force at the end of the beam. The function Φ_i, ψ_i, $i = 1,2$ are the functions defined on the domain shown in Fig. 4.2b.

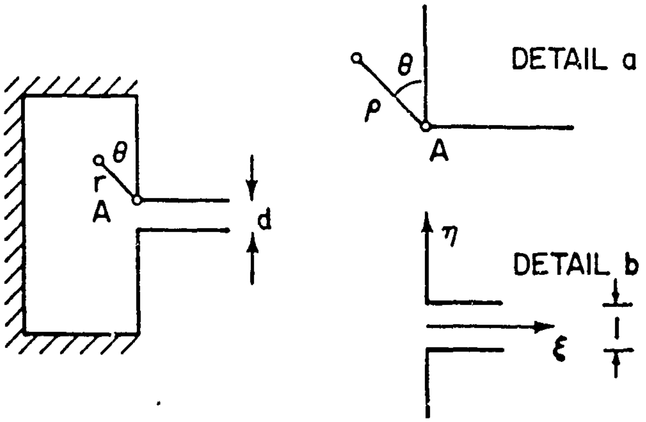


Fig. 4.2. The solution on the zoomed domain.

Coefficient M and T are analogous to the stress intensity factors introduced earlier. The zooming principles can be used in many cases when the corner singularity interferes as in the case shown in Fig. 4.2. For more details we refer to [6].

4.3. The problem in 3 dimensions

In 3 dimensions we will consider a polyhedron instead of a polygon and the problem becomes more complex. Once more we will consider the case of isotropic homogeneous maeterial. *Along the edges* the solution is singular in the direction which is perpendicular to the edges and is smooth along the edges. Assuming that the (straight) edge is along the axis z, the singular terms are of the form

$$\left.\begin{array}{l} u(x,y,z) \\ v(x,y,z) \\ w(x,y,z) \end{array}\right\} = c_i(z) r^{\lambda_i} \left\{\begin{array}{l} \varphi_i(\theta) \\ \eta_i(\theta) \\ \zeta_i(\theta) \end{array}\right.$$

where (r,θ,z) are the cylindrical coordinates. Functions $\varphi_i, \psi_i, \zeta_i$ are smooth in θ. There are, in contrast to the two dimensional case, two kinds of singular function. For the first kind coefficients λ_i and func'ions φ_i, ψ_i are as in 2 dimensions and $\zeta_i = 0$ (they are sometimes called bending singularities). In the second kind we have $\varphi_i = \psi_i = 0$ (and $\xi_i \neq 0$) (they are sometimes called torsion singularities). For details we refer to [15], [22]. The function $c_i(z)$ are the stress *intensity functions* which are smooth in z (except the neighborhood of the verticles).

In addition to the edge type singularity, we have a *vertex singularity*. Here the singular terms have the form

$$\left.\begin{array}{l} u(x,y,z) \\ v(x,y,z) \\ w(x,y,z) \end{array}\right\} = c_i \mathbb{R}^{\Lambda^{(i)}} \left\{\begin{array}{l} \bar{\varphi}_i(\theta,\Xi) \\ \bar{\psi}_i(\theta,\Xi) \\ \bar{\zeta}_i(\theta,\Xi) \end{array}\right.$$

where $(\mathbb{R},\theta,\Xi)$ are the spherical coordinates. Functions $\varphi_i, \psi_i, \zeta_i$ have singular behavior in the neighborhood of (θ_s, ∂_s) being the coordinates of the edges. For details once more see [15], [22].

The coefficient $\Lambda^{(i)}$, and functions $\bar{\varphi}_i, \bar{\psi}_i, \bar{\zeta}_i$ depend on the geometry, boundary conditions and material properties but not on the solution. The coefficients c_i are the analog to the stress intensity factors and depend (except special case) *globally* on the solution. The relation between λ_i and $\Lambda^{(i)}$ governs the singular behavior of the edge intensity factor function in the neighborhood of the vertex. The unboundedness of the stresses in the neighborhood of the vertex take place when $\Lambda^{(1)} < 1$ provided the coefficient $c_1 \neq 0$. Because of *global* dependence of c_1 on the solution, $c_1 \neq 0$ practically always (except possible in special case of symmetries). Nevertheless c_1 can be small and large stresses can be confined to a small area. Then the usual finite element solution could indicate completely wrong behavior. *Hence computation of* c_i *is a necessity to obtain reliable results.*

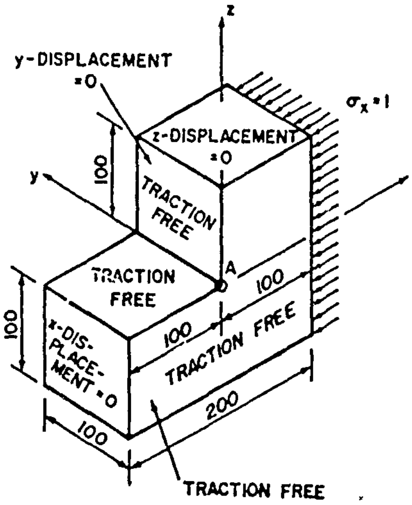


Fig.4.3. The 3 dimensional domain

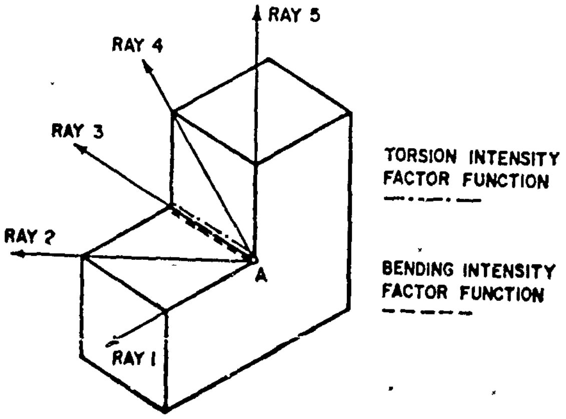
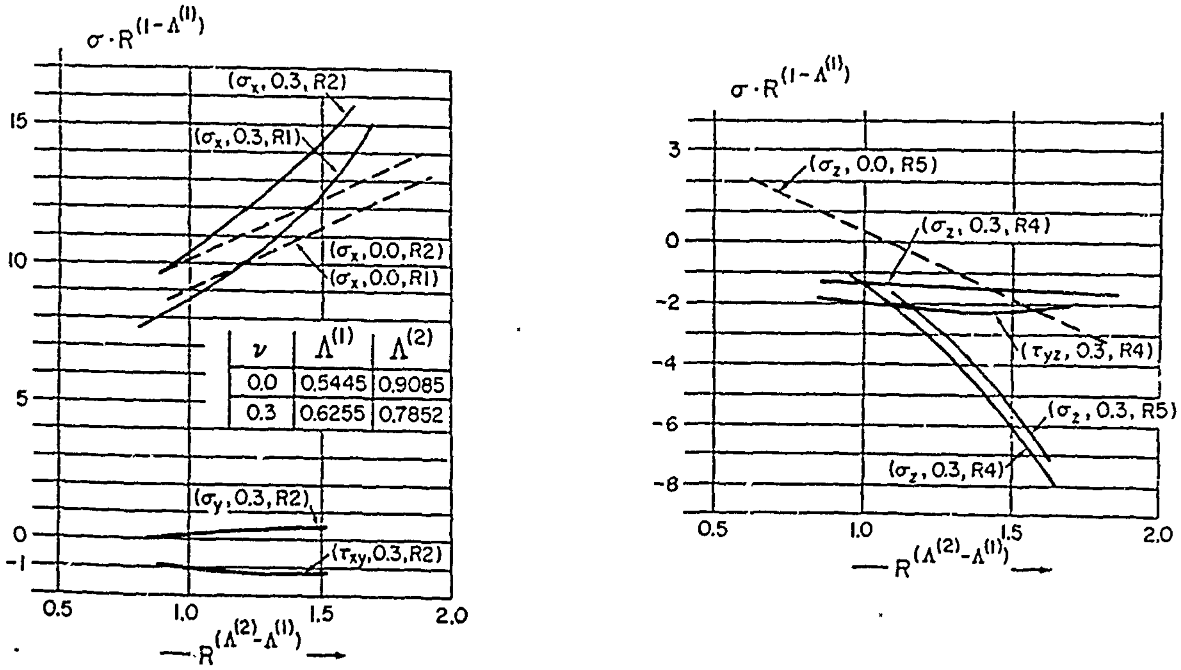

RAY 5
RAY 4
RAY 3
RAY 2
A
RAY 1
TORSION INTENSITY FACTOR FUNCTION
—·—·—
BENDING INTENSITY FACTOR FUNCTION
– – – –

Fig.4.4. Location of the rays

Let us show now two typical examples. Fig.4.3 shows a three dimensional domain with imposed boundary conditions. Along the marked edge shown in Fig. 4.4 the stress intensity factor functions are present. In the neighborhood of A solution has vertex type of singularity. In Fig.4.4 we show the rays where the stresses are depicted. The first two coefficients Λ_i are given in Fig.4.5a for $\nu = 0.0$ and $\nu = 0.3$. The stresses on the rays will have then the form

$$\sigma = C_1 R^{\Lambda^{(1)}-1} + C_2 R^{\Lambda^{(2)}-1} + \text{higher order terms.}$$



ν	$\Lambda^{(1)}$	$\Lambda^{(2)}$
0.0	0.5445	0.9085
0.3	0.6255	0.7852

Fig.4.5. The behavior of stresses on the rays.

and

$$\sigma R^{1-\Lambda^{(1)}} = C_1 + C_2 R^{(\Lambda^{(2)}-\Lambda^{(1)})} + \text{higher order terms.}$$

Hence in the scale $\sigma R^{1-\Lambda^{(1)}} \times R^{(\Lambda^{(2)}-\Lambda^{(1)})}$ the behavior is linear for small

R. C_1 and C_2 relate to the need to compute two *stress* singular functions. Fig.4.5 ab depict the stresses on different rays. The Fig.4.5ab show well also the scales where the large stresses will appear (which is of the order of 1/100 of the thickness). We also see that the value of ν does not signifintly influence the behavior of the solution. Nevertheless this is not always the case as we will see in the next example.

The second problem (which was suggested by K.J. Bathe, (see also [9]) is depicted in Fig. 4.6.

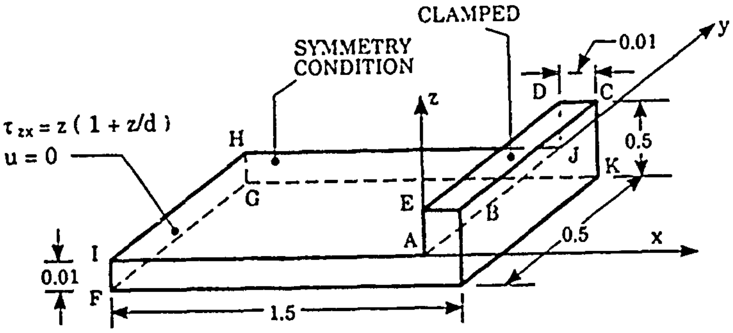


Fig. 4.6 The Bathe's problem

The boundary conditions are shown in the figure. When nothing is explicitly stated then the associated tractions are zero. Let us consider now the behavior of the solutions at the edge I-A: for $\nu = 0.0$ and $\nu = 0.3$. The values of the constants $\Lambda^{[1]}$ are the same as in the previous example. Fig.4.7 shows now the stress σ_x at the edge I-A. We see here drastic difference between the stress behavior for $\nu = 0.0$ and $\nu = 0.3$. The standard finite element analysis will lead to the conclusion that σ_x is bounded on the edge I-A for $\nu = 0.3$ while for $\nu = 0.0$ is large. This conclusion is of course completely wrong. Let us mention that the strength of material solution predicts $\sigma_x = -280$ for $x = -d$ and $z = 0$. This value is approximately achieved for $\nu = 0.0$ for any y (the problem here is

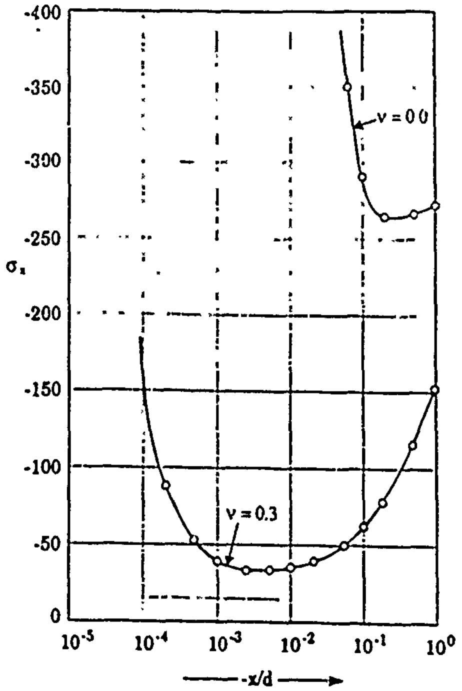


Fig. 4.7 The stress σ_x on I-A for $\nu = 0.0$ and $\nu = 0.3$

y-independent). For $\nu = 0.3$, it is approximately achieved for $y > 7d$[*]. We see that for a reliable conclusion about the stresses in the neighborhood of the corners and edges the computation of the stress intensity factor *is essential* and any program should always have to be able to provide them (STRIPE provides them). For more about the problem of reliable computations in solid mechanics and detailed analysis of Bathe problem and engineering we refer to [1].

5. The plate problem

The plate (and shell) problem is a basic problem in engineering. As the *basic mathematical problem* we will understand the three dimensional problem of elasticity on the thin domain

[*] Analysis of the problems of this section has been made by programe STRIPE by Dr. B. Andersson. For more see also [1].

$$\Omega = \{x,y,z \mid (x,y) \in \omega,\ |z| < d/2\}.$$

We will assume isotropic homogeneous material.

5.1 The problem of the derivation of the simplified formulation.

There are very many formulations of the plate problems. For the survey see [20] [23]. The major simplified formulations are the Kirchhoff [K] and Reissner-Mindlin (RM) formulations. Many results describe the asymptotic behavior of the solution of 3 dimensional (basic) formulation as $d \to 0$. These results show that (for example in the energy norm) the 3D and Reissner-Mindlin solutions converge to the Kirchhoff solution as $d \to 0$, see eg. [12], [19]. For a detailed study of the asymptotic behavior of Reissner-Mindlin problem we refer to [2].

Further there are generalized models based on the projection in the energy on the space of functions of the form (Kantorovich method)

$$(5.1) \qquad U(x,y,z) = \sum_{k=0}^{n} \varphi_k(x,y)z^k$$

$$V(x,y,z) = \sum_{k=0}^{n} \psi_k(x,y)z^k$$

$$W(x,y,z) = \sum_{k=0}^{m} \xi_k(x,y)z^k.$$

In the case $\nu = 0$, the case $n = 1$ with $\varphi_0 = \psi_0 = 0$ and $m = 0$ leads to RM model. For $\nu > 0$, $n = 1$, $m = 2$, $\varphi_0 = \psi_0 = \xi_1 = 0$ leads to RM model with singular perturbations. Usually $m = n+1$ is taken, which guarantees the proper asymptotic rate of convergence. The model based (5.1) will be called n-m model. In general n and m can be different in different parts of ω. For $n,m \to \infty$ the solution of (n-m) model converges to 3 dimensional solution (in the energy norm).

The form (5.1) leads to a hierarchic family of models. It depends on particular choice of the function in z. In (5.1) polynomials have been used. This choice is optimal in an asymptotic way when $d \to 0$ (see e.g. [12], [26]). Other optimal choices can be considered too [24].

The error of the various models have to be judged in the relation to *the* 3D *solution* and which data are of interest.

5.2 The problem of the rhombic simply supported plate.

Let us consider the plate shown in Fig.5.1. The simple support can be formulated as

a) hard simple support

b) soft simple support

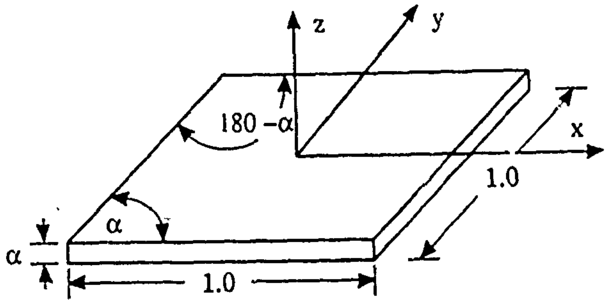


Fig. 5.1 The rhombic plate

In the case a) we assume on the lateral sides $w = 0$ and $u_t = 0$ where by u_t we denote the displacement in the directions of the tangent to the boundary of ω (and hence $T_t \neq 0$), u_n is free (i.e. $T_n = 0$).

In the case b) the only constraint on the lateral side is $w = 0$ (and hence $T_n = T_t = 0$). The K-model cannot distinguish between these two supports. Few problems now appear

i) How much do the solutions for the two models of support differ?

ii) Which support does the K-model describe?

iii) How accurate (with respect to the 3D model) are the K and RM models?

Answers to these questions depend strongly on how we measure the error. It is well known that the major difference between K and RM model is in *the boundary layer*. For a detailed analysis in the case when the boundary of ω is smooth we refer to [2].

If we take the energy norm measure of the error and the load uniform on the upper side of the rhombic plate the relative error in % for the K-model, and the $m = n = 2$ model (which leads to the slightly smaller error the RM model) and for the *soft simple support* is given in Table 5.1 (3D solution is taken as exact).

Table 5.1 The relative energy norm error of K and $m = m = 2$ model for soft simple support in %

α	$\nu = 0.0$				$\nu = 0.3$			
	d = 0.1		d = 0.01		d = 0.1		d = 0.01	
	K	(2,2)	K	(2,2)	K	(2,2)	K	(2,2)
90°	39.56	12.57	11.87	3.50	34.52	11.18	9.88	2.94
80°	39.91	12.59	12.23	3.57				
60°	42.24	12.72	15.46	4.14				
40°	45.43	13.60	20.50	4.24				
30°	48.27	15.41	22.66	4.34	44.68	15.03	18.91	3.68

Table 5.2 shows for $\alpha = 90$ and $\nu = 0$ the error for the hard support (the 3D solution with hard support is taken as exact). The error of the 3D solution with the hard support with respect to the soft support of 3D formulation is 34.7% for $d = 0.1$ and 11.7% for $d = 0.01$ For more see [19].

Table 5.2 The relative enery norm error for the K and (2.2) model for hard support in %.

Model	d = 0.1	d = 0.01
K	20.31	2.03
(2.2)	8.22	0.68

The difference between models eg. K, RM, n-m and 3D model is largest in the boundary layer. It can be in fact very large.

To illustrate this, we consider the square plate $(\alpha = 90,\ \nu = 0.0)$. Let Q_{xz} and Q_{yz} be the shear forces on the line $x = 0.5,\ 0 < y < 0.5$ computed from the 3 dimensional solution for $d = 0.01$ and soft support. They are shown in Fig.5.2ab. Realizing that the K-model leads to $Q_{yz} = 0$, we see that K-model is unreliable for these data of interest.

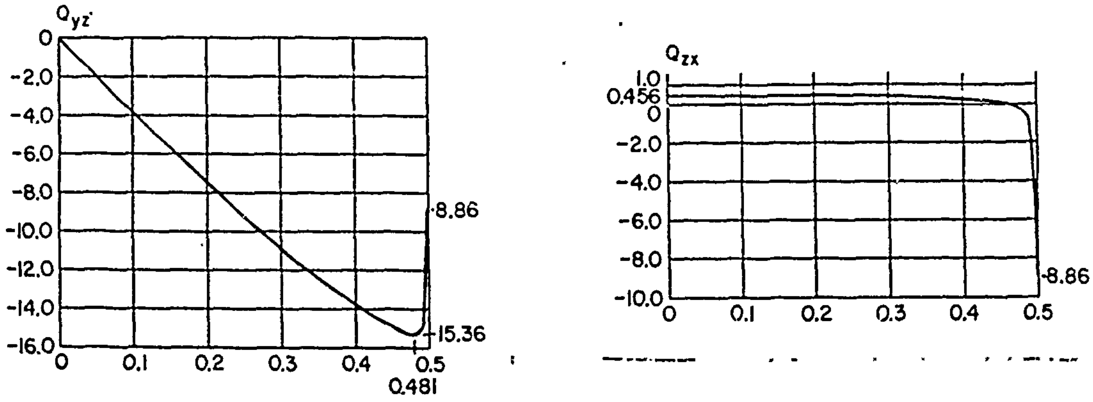
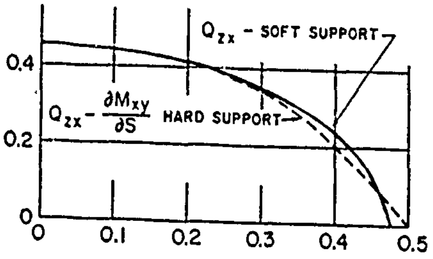


Fig.5.2 The shear force Q_{xz} and Q_{yz} at $x=0.5$ and $0<y<0.05$

The K-model approximates relatively well the *hard support* but not the soft one. In the larger distance from the boundary, the K-model is usually usable also for the soft support. Although the K-model approximates the hard support the usual approach is to modify the reaction Q_{xz} by the derivative of the twist moment to get the soft support reaction. This leads to a reasonably good approximation of Q_{xz}. In Fig.5.3 Q_{xz} for soft and hard support with twist moment adjustment is shown. Nevertheless no adjustment of K-model could give reasonable values of Q_{yz}.



5.3 The reaction Q_{xz}.

There is a significant theoretical difference between hard and soft sample support.

In [6] we have analyzed the behavior of the solutions on a regular n-eck polygons ω_n inscribed in the unit circle. We have shown that for hard simple support the solution in ω_n converge to a solution on the circle S but which paradoxically _is not_ the solution on S for the hard simple support. This happens for K, RM and 3D model. In contrast, in the case of the soft support, such paradox does not occur. This shows that the mathematical problem of the hard support has a property which we would not expect in "reality" and hence an increased precaution has to be given when nard support model is used.

5.2 The problem of the solution singularity of the plate models.

Let us once more consider the square plate ($\alpha = 90^\circ$), d = 0.01, $\nu = 0.3$ such that the sides $0.5 < x < 0.5$, $y = \pm 0.5$ *are clamped* (u, v, w = 0) and other two sides *are free.* Then in the neighborhood of the vertices the solution is singular. In [11], [27] the singular behavior of the RM solution is analyzed. This analysis shows that RM solution has two different characters of the singularities. Denote by (r,θ) the polar coordinates with center in $A = (0.5,0.5)$. Then for $r \ll d$ the solution for example the stresses or moments have the form

$$M_x \cong \sigma_x \cong Cr^{\lambda_{RM}}\varphi_{RM}(\theta)$$

while for $r \approx d$, the singularity is as for the Kirchhoff model

$$\sigma_x \cong Cr^{\lambda_K}\varphi_K(\theta)$$

Between these two areas there is a transition domain. The exponents λ_{RM}, λ_x satisfy some transcendental equations. In our case

$$\lambda_{RM} = -\ 0.241$$
$$\lambda_K = +\ 0.0686 + i0.438$$

(i.e. the stress of K-solution is oscillating). In Fig. 5.4 we show the stress σ_x at the diagonal of the plate in log - log scale. We see clearly that both types of singularities occur. Other stresses show similar behavior.

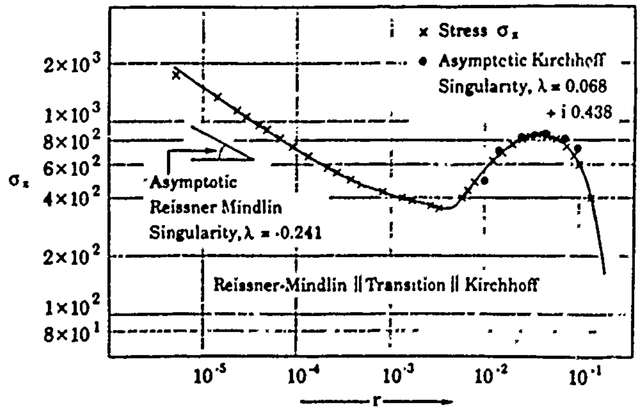


Fig. 5.4 Stress σ_x of the RM model.

Finally we can compute the character of the moments computed from the 3 dimensional solution. Here we can show that

$$M_x \approx Cr^{\lambda_{3D}}\varphi_{3D}(\theta)$$

and in our case

$$\lambda_{3D} = -0.289$$

We see that the corner behavior is different for these 3 models. This difference strongly depends on the geometry of the plate and boundary conditions. For more we refer to [8].

6. The problem of nonlinear elasticity

The nonlinear formulation in the theory of elasticity stems from

a) nonlinear geometry as large displacements, stresses, etc.

b) nonlinear constitutive law.

Here we will address some questions related *to the elasticity* assuming static behavior where effects of velocity etc. can be neglected. In *one dimensional case*, given the strain $\varepsilon(t)$ $-\infty < t < \infty$ the constitutive law leads to the stress response $\sigma(t)$

$$\sigma = A\varepsilon \tag{6.1}$$

where A is an operator mapping the strains into the space of stresses.

In the 3 dimensional case, ε and σ are the strain and stress tensors, respectively.

Usually the constitutive law in 2 and 3 dimensions is derived from the one dimensional law by applying various principles as Mises, Tresca, Huber-Hencky, etc. In one dimension many laws were proposed, see e.g. [28]. The basic laws are kinematic, isotropic hardening and others. Recently the formulation by Chaboche [13] has become popular.

Mathematically it is important that the constitutive law is such that it satisfies conditions which guarantee the desirable properties of the mathematical problem of elasticity where it is used. It is of course also important that there is not a large difference between observed and predicted response (based on the constitutive law used).

6.1 Experimental results

Results of an extensive one dimensional, experimental analysis with the aluminum alloy 5454 in the H32 condition are reported in [18]. This alloy is produced (under the same commercial mark) by different manufacturers and is widely used in engineering.

The analysis in [18] is based on the fact that in engineering the material is taken from the warehouse and at best the experiments for selection of the proper constitutive law can be made on samples only (statistical approach). Hence 84 samples have been taken and analyzed. Among others, the main questions were related to

a) reproducibility of the response

b) selection of the constitutive law.

Two classes of the strain were considered

i) the cyclic periodic strain (which is usually used in materiai science).

ii) random strain which is more realistic in applications.

The main results can be broadly characterized as follows

a) The reproducibility factor Q_R for the random strain is of order ≈10-15% where

$$Q_R = \frac{\max_t |A(t) - B(t)|}{\max_t \left| \frac{A(t) + B(t)}{2} \right|}$$

Here $A(t)$ and $B(t)$ is the stress response of two different sample to the same *random* strain. For *cyclic* load the factor Q_c is of order 7-10%.

b) For every sample and particular constitutive law mostly used in practice (as Chaboche, kinematic, Mroz, etc.) the constants for the best fit were computed. Then the average value of these (84 samples) were computed and using these constants the constitutive law the factors C_R (resp. C_c) were analogous to Q_R(resp Q_c) i.e. we define

$$C_R = \frac{\max\limits_t |A(t) - \bar{B}(t)|}{\max\limits_t \left|\frac{A(t) + \bar{B}(t)}{2}\right|} .$$

Here $A(t)$ is the response for a sample and $\bar{B}(t)$ is the predicted response based on the average constants. For the best law (one of them is Chaboche we get $C_R \approx 22\text{-}25\%$ and $C_c \approx 16\text{-}18\%$.

For the best fit of one sample we get $C_R \approx 8\%$.

If the set of averaging is small we can get $C_R > 30\%$.

For some laws (in standard use in FE codes), $C_R > 40 - 50\%$.

In the Table 6.1 we show the $\|\cdot\|_{L_\infty}$ $\|\cdot\|_{L_2}$ norm (in psi) and relative error.

		$\|A-B\|_{L_\infty}$	$\|A-B\|_{L_2}$	$\frac{\|A-B\|_{L_\infty}}{\|(A-B)/2\|_{L_2}}$	$\frac{\|A-B\|_{L_2}}{\|(A-B)/2\|_{L_2}}$
fd	fq	5322	2334	14.4%	12.9%
fd	Chabache	8346	2654	22.0%	13.5%
fd	Kinematic	11850	3475	32.8%	17.6%

(fd, fq is the label of the sample, Chaboche and Kinematic means response obtained by the Chaboche resp. kinematic law). We mention that for computational purpose, the norm $\|\cdot\|_{L_\infty}$ is essential (and not $\|\cdot\|_{L_2}$).

As an illustration we show in Fig. 6.1, 6.2, 6.3, the value for A-B (two samples) and A-$\bar{B}$ for character and kinematic law using average constants, for the random strain.

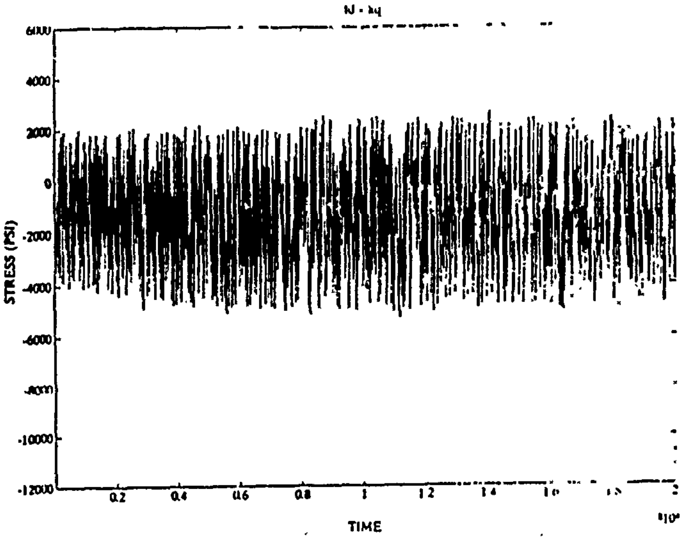


Fig.6.1 The difference between two samples

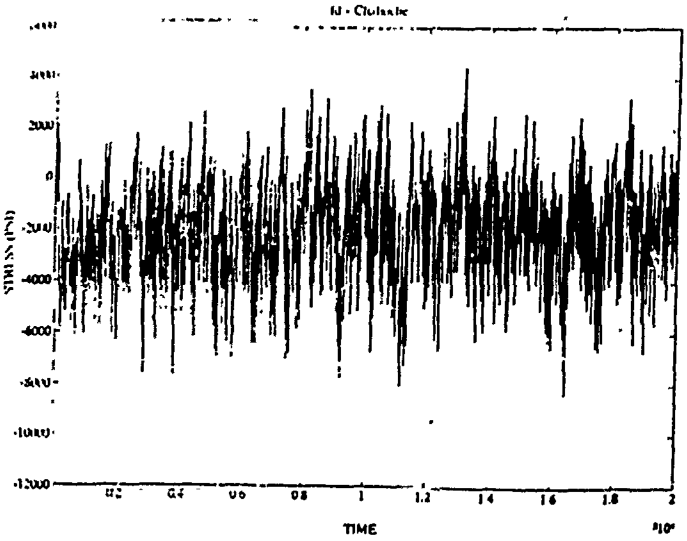


Fig.6.2 The difference between sample and Chaboche law

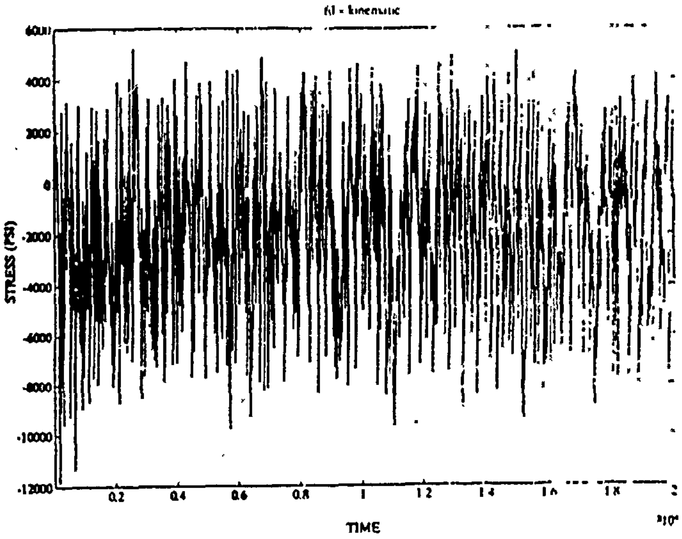


Fig.6.3 The difference between sample and kinematic law.

[18] analyzes only the one dimensional problem. It is possible to expect that in 2 resp. 3 dimensional setting the factors will be larger.

The analysis made in [18] indicates

a) It is necessary to analyze the reliability of constitutive laws derived from statistical sampling.

b) It is necessary to analyze random and not cyclic strains.

c) It is highly desirable to develop a mathematical theory for determining the constitutive law based on the (infinitely dimensional) identification problems and to develop a strategies for optimal selections of strains for experiments. This is especially important for 2 and 3 dimensional settings.

d) It seems that the usual elasticity formulation and computation based on the "average" constitutive law cannot give reliable results and other approaches such as bracketting have to be developed, see also here [10].

6.2 Mathematical formulation of Chaboche law.

We have seen in the Section 6.1 that the Chaboche law is one of the laws which fits best the data for the single sample. Therefore we will discuss it here in more detail.

Although in [13] the law is formulated in an incremental way related to mechanical interpretation, it can be cast into a system of ODE for the stress and *two* (internal) parameter functions, $(\sigma(t), \chi(t), R(t))$ for given strain $\varepsilon(t)$. In what follows we denote $\dot{\varepsilon}(t) = \frac{d\varepsilon}{dt}$ etc. The Chaboche law is characterized by 6 constants.

We have

$$\dot{\sigma} = E\dot{\varepsilon}, \qquad \sigma(0) = \chi(0) = R(0) = 0$$

$$\dot{\chi} = 0, \qquad \varepsilon_h(0) = \varepsilon_y, \ \varepsilon_\ell(0) = -\varepsilon_y$$

(6.2)
$$\begin{aligned} \dot{R} &= 0 \\ \dot{\varepsilon}_h &= 0 \\ \dot{\varepsilon}_\ell &= 0 \end{aligned}$$

for all $t \in \mathcal{E}$, where

$$\mathcal{E} = \{t \,|\, \varepsilon_\ell(t) < \varepsilon(t) < \varepsilon_h(t), \text{ or } \varepsilon(t) = \varepsilon_h(t) \text{ and } \mathring{\varepsilon} \le 0 \text{ or } \varepsilon_\ell(t) = \varepsilon(t) \text{ and } \dot{\varepsilon} \ge 0\}$$

$$\dot{\sigma}(t) = \frac{E[c(a-x(t)) + b(Q-R(t))]}{c(a-\chi(t)) + b(Q-R(t)) + E}\,\dot{\varepsilon}(t)$$

(6.3)
$$\dot{\chi}(t) = \frac{E[c(a-x(t))}{c(a-\chi(t)) + b(Q-R(t)) + E}\,\dot{\varepsilon}(t)$$

$$\dot{R}(t) = \frac{Eb(Q-R(t))}{c(a-\chi(t)) + b(Q-R(t)) + E}\,\dot{\varepsilon}(t)$$

$$\dot{\varepsilon}_h = \dot{\varepsilon}$$

$$\dot{\varepsilon}_\ell = \dot{\varepsilon} - 2\frac{R}{E}$$

for all $t \in P_+$ where

$$P_+ = \{t \,|\, \mathring{\varepsilon} > 0 \quad \text{and} \quad \varepsilon = \varepsilon_h\}$$

and

$$\dot{\sigma}(t) = \frac{E[c'\,\imath{-}x(t)) + b(Q-R(t))}{c(a+\chi(t)) + b(Q-R(t)) + E} \tag{6.4}$$

$$\dot{R}(t) = \frac{tb(Q-R(t))}{c(a+\chi(t)) + b(Q-R(t)) + E}$$

$$\dot{\varepsilon}_h = \dot{\varepsilon} - 2\frac{R}{E}$$

$$\dot{\varepsilon}_\ell = \dot{\varepsilon}$$

for all $t \in P_-$, where

$$P_- = \{t \,|\, \mathring{\varepsilon} < 0 \quad \text{and} \quad \varepsilon = \varepsilon_\ell\}$$

Chaboche model is characterized by 6 constants with a physical interpretation.

a: Kinematic coefficient

c: Kinematic exponent

Q: isotropic exponent

b: isotropic exponent

g: yield strain

E: elastic modulus

The Chaboche law as formulated can be generalized into 2D and 3D formulations. Nevertheless when this is used in the nonlinear elasticity equation problem, some desirable mathematical properties of the problems (as for example, the existence of the solution) are not quaranteed. Hence another formulation which approximate well the Chaboche law and leads to the desirable properties of the mathematical formulation should be used.

6.3 A proper mathematical formulation of the constitutive law.

Let us outline now here principles and family of constitutive laws (called gauge method) which guarantee good properties of the mathematical problems based on them. There are two basic (sufficient) conditions for it.

a) Existence and convexity of the yield surface

b) the normality condition

(These conditions are related to the Druckers postulates).

Let $\underline{\alpha} \in \mathbb{R}^m$ be the set of internal parameters, $\underline{\alpha} \subset A \subset \mathbb{R}^m$, A being a convex set in $\mathbb{R}^m$. Set further $\sigma \in \mathbb{R}^3$ (for a two dimensional problem).

Now we will formulate the law with the help of the yield function. To this end let $F(\sigma,\alpha)$, $F : \mathbb{R}^3 \times A \longrightarrow R$ be given so that

(6.8a) F is convex and C^1

(6.8b) $F(0,0) = 0$

(6.8c) There exist constants γ, Γ such that $0 < \gamma < |\delta_\alpha F|\,|\delta_\alpha F| < \Gamma$

uniformly on the set $\{(\sigma,\alpha)|F(\sigma,\alpha) = z_0\}$ for some z_0.

Then

(6.9)
$$\dot{\sigma} = D\dot{\varepsilon} \quad \text{if} \quad t \in \mathcal{E}$$
$$\dot{\alpha} = 0$$

(6.10)
$$\dot{\sigma} = \left[D - \frac{D\delta_\sigma F(\delta_\sigma F)^T D}{(\delta_\alpha F)^T(\delta_\alpha F) + (\delta_\sigma F)^T(\delta_\sigma F)} \right] \dot{\varepsilon}$$

$$\dot{\alpha} = -\left((\delta_\alpha F)^T - (\delta_\alpha F)\right)^{-1}\left((\delta_\sigma F)^T\dot{\sigma}\right)\delta_\alpha F \quad \text{if} \quad t \in \mathcal{P}$$

where

$$\mathcal{E} = \{t|F(\sigma,\alpha) < z_0 \quad \text{or}$$

$$F(\alpha,\alpha) = z_0 \quad \text{and} \quad (\delta_\alpha F)^T\dot{\sigma} \leq 0\}$$

$$\mathcal{P} = \{t \mid F(\sigma,\alpha) = z_0 \quad \text{and} \quad (\delta_\sigma F)^T \dot{\sigma} \geq 0\}.$$

The Chaboche model could be cast approximately in the above frame using

$$F(\sigma,\alpha,\beta) = [\max(F_1(\sigma,\alpha.\beta),\ F_2(\sigma,\alpha.\beta)]^*$$

where

$$F_1(\sigma,\alpha.\beta) = a_1(\alpha-\alpha_1)^2 + a_2(\alpha-\alpha_1) + a_3(\beta-\beta_1)^2 + a_4(\beta-\beta_1) + (\sigma-\sigma_1) + \zeta_1 \tag{6.10a}$$

$$F_2(\sigma,\alpha.\beta) = b_1(\alpha-\alpha_2)^2 + b_2(\alpha-\alpha_2) + b_3(\beta-\beta_1)^2 + b_4(\beta-\beta_2) - (\sigma-\sigma_2) + \zeta_2 \tag{6.10b}$$

and [ ]* the smoothing the operator in the neighborhood of the manifold $F_1(\sigma,\alpha,\beta) = F_2(\sigma,\alpha,\beta)$

Fig. 6.4a shows the relation between ε and σ for cyclic (sinusoidal) strain with 50 reversals (25 periods) for the constant computed out of the experimental data (averages of Chaboche constants).

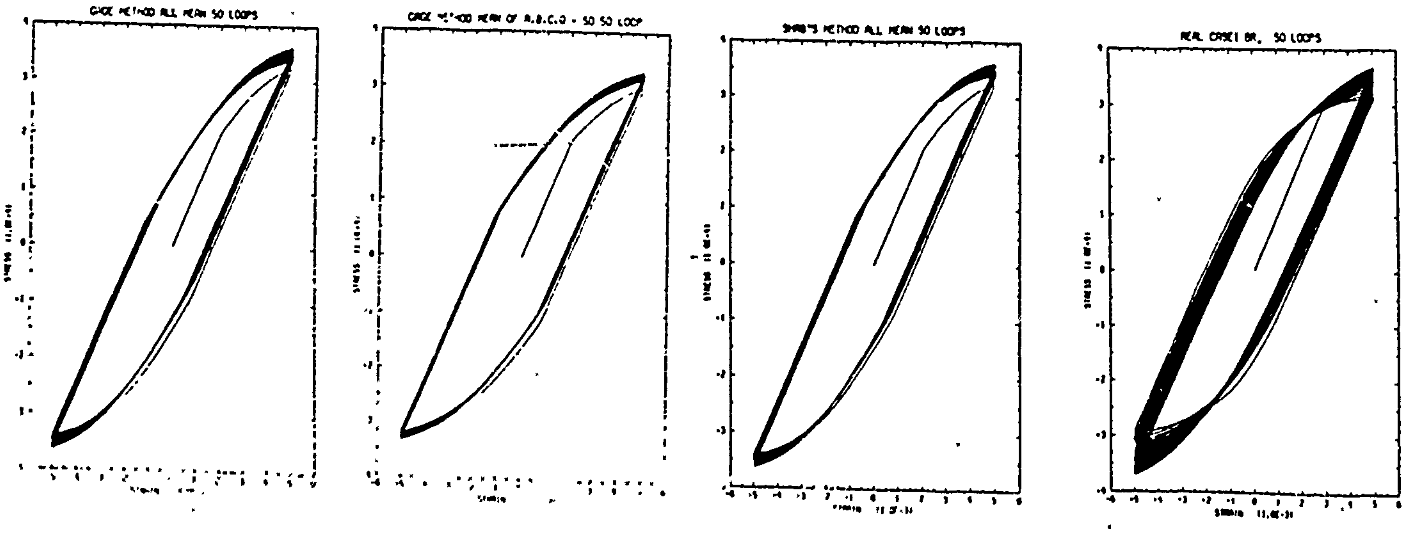


Fig.6.4 The relation between the strain ε and the stress for the law based on (6.10a,b).

Fig. 6.4b shows the results where, for the constant, we have taken the mean minus standard deviation (let us mention that the correlation for these coefficients (see [18]) are of order 0.2). Fig. 6.4c shows the results when computed from the Chaboche law (see section 6.2) and Fig.6.4d shows the experimental results for one sample.

We see that the results from the original Chaboche law are well approximated. In all these data we have assumed that the initial data which depend on past history have been known. In reality they are not known which further increases the uncertainties in the available information. The derivation of 2 and 3 dimensional constitutive law leads to still more uncertainties because not enough experiments could be made. In [10] the 2 dimensional constitutive law was derived as the limit of the frame made out of the bars, analogous Cauchy's derivation of linear elasticity.

We have seen that the computation of the problem of elasticity on the assumption of the knowledge of constitutive law without respecting the uncertainties leads to unreliable results.

7. A posteriori error analysis of the model.

It is essential to make a posteriori analysis of the error of the solution of the simpliefied problem when only the data from this simplified model are used. This can often be made by two sided energy estimates. For details see e.g. [17].

7.1. Estimate of the error of geometry idealization.

Consider the problem on Ω_r shown in Fig. 7.1.

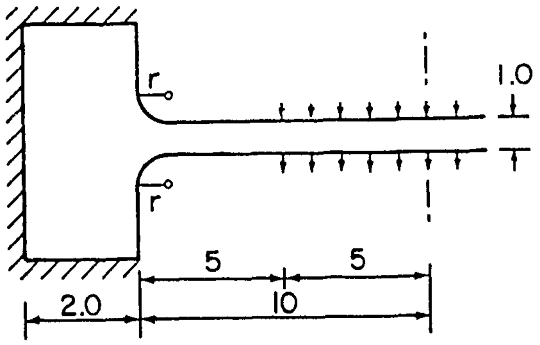


Fig. 7.1. Scheme of the problem with perturbed boundary.

Let the basic mathematical problem be the linear elasticity on Ω_r and the simplified problem is the problem on Ω_0 using $E = 1$, $\nu = 0.3$ (and $r = 0$). Then by finite element solution we get two stress intensity factors (see Section 4) $c_1 = 0.2157\ 10^2$, $c_2 = 0.5929\ 10^2$. Then the estimate using c_1, c_2 (and the form of the singular functions) allows to compute the upper estimate in the energy norm on Ω_0 of the difference between the solutions on Ω_r and Ω_0 (see [17]). Table 7.1 gives the results together with the true error obtained by the solution on Ω_r.

Table 7.1 The estimate and true error of the geometry ideal

r	Estimate	True error
0.1	19.0%	13.2%
0.01	3.5%	3.0%

We see good effectiveness of the estimate.

7.2 Estimates of the linearization

Let us consider once more the problem shown on Fig.7.1 with $r = 0$. If the simplified problem will be understood as the linear problem the strains and stresses are infinite (see Section 4). If we would consider as the basic problem the nonlinear problem, the nonlinearity will occur in the neighborhood of the corner. We will assume the Hencky model of the nonlinear elasticity with the governing function

$$\varphi(\zeta) = \begin{cases} \zeta & \text{for } 0 \le \zeta \le \zeta_0 \\ \beta\zeta + \frac{1-\beta}{y}\zeta_0^{1-\gamma}\zeta^{\gamma} - \frac{(1-\gamma((1-\beta)}{\gamma}\zeta_0 & \text{for } \zeta > \zeta_0 \end{cases}$$

with $\gamma \in \left[\frac{0.5-\beta}{1-\beta}, 1\right)$.

Functions $\varphi(\zeta)$ is the function describing the nonlinearity of the material.

Let us consider the problem depicted in Fig.7.1 with $E = 10^6$, $\nu = 0.3$. Table 7.2 shows the upper estimate of the relative error mentioned in the energy norm for various values of β, γ, ζ_0. For more details we refer to [17].

$\zeta_0 = 0.01$			$\zeta_0 = 0.001$		
β	γ	Estimate of relative error	β	γ	Estimate of relative error
0.1	0.501	2.76 10^{-4}%	0.9	0.8	2.05 10^{-5}%
0.1	0.5001	2.79 10^{-4}%	0.9	0.5	4.06 10^{-5}%
0.1	0.50001	2.79 10^{-4}%	0.9	0.02	5.36 10^{-5}%
0.01	0.501	9.56 10^{-4}%	0.5	0.8	1.19 10^{-4}%
0.01	0.5001	9.62 10^{-4}%	0.5	0.5	2.55 10^{-4}%
0.01	0.50001	9.63 10^{-4}%	0.5	0.2	3.89 10^{-4}%
0.001	0.501	3.03 10^{-3}%	0.3	0.3	1.84 10^{-4}%
0.001	0.5001	3.07 10^{-3}%	0.3	0.5	4.32 10^{-4}%
0.001	0.50001	3.08 10^{-3}%			

8. Conclusions

We have discussed various aspects of reliability without trying to define more precisely what does it mean. We have seen that the aim is to get desired data with an assessment of their accuracy. More precisely we are aiming to get the quantitative bracketts in which the "true" results are. These bracketts then express the uncertainty of the results caused by uncertainties in the input data, the (simplified) formulation, the discretization etc.

We can now roughly define the reliability of engineering computations in the *relation to reality.*

The computational results furnished with the bracketts are physically reliable if the physically observed results are in the provided bracketts.

Analogously, (more precisely) we can define the reliability of the computational results in *the relation of mathematical analysis.*

The computational results furnished with the bracketts are mathematically reliable if the exact data of the basic mathematical problems are in the provided bracketts.

We have seen that the reliability is related to the data of interest and the definitions what is meant by accuracy (e.g. particular norms) etc.

We have also seen that the mathematical formulation has to be closely related to the engineering analysis and experimentation. Without it, the "physical" reliability is impossible to expect.

The mathematical reliability, i.e. the comparison of the obtained results with the exact data stemming from the *basic mathematical problem* is always (at least in principle) possible. This comparison and its bracketting is then the main goal of the (mathematical) computational analysis.

The reliability of the computational analyses has many features and bring out many unsolved problems. Nevertheless there are already today many ways to get at least partial *quant:tative* insight into reliability of computed results.

REFERENCES

1. B. Andersson, I. Babuška, U. Falk T. Petersdorff, Accurate and Reliable Computation of Complete Solutions of Equations of Linear Elastomechanics on Three-dimensional Domains, to appear.

2. D.N. Arnold, R. Falk, Edge Effects in the Reissner-Mindlin Plate Theory, Preprint, Symposium on Analytical and Computational Methods for Shells, December 1989, San Franciso, to appear in Proceedings of Am. Soc. of Mech. Engineers.

3. I. Babuška, Uncertainities in Engineering Design, Mathematical Theory and Numerical Experience, in J.A. Bennett and M.E. Botkin, eds. The Optimum Shape pp.171-197, Plenum Press, New York 1986.

4. I. Babuška, J. Pitkäranta, The Plate Paradox for Hard and Simple Support. SIAM J. Math. Anal. 1990.

5. I. Babuška, A. Miller, The Postprocessing Approach in Finite Element Method, Int. J. Num. Meth. Engrg. 20 (1984), 1085-1109, 111-1129, 2311-2324.

6. I. Babuška, T. Petersdorff, Boundary Layers in Beams with Various Boundary Coditions, to appear.

7. I. Babuška, T. Scapolla, Benchmark computation and performance evaluation for a Rhombic Plate Bending Problem, Int. J. Num. Meth. Engrg. 28(1989), 155-179.

8. I. Babuška, L.Li Corner singularities of the solution of various plate models, to appear.

9. K.J. Bathe, N.S. Lee, M. Bucalem, On the use of hierarchical models in engineering analysis, to appear.

10. E. Bonnetier, Mathematical Treatment of the Uncertainties appearing in the Formulation of some Models for Plasticity. Ph.D. Thesis (1988), University of Maryland, College Park, MD 20742, USA.

11. W.S. Burton, G.B. Sinclair, On the Singularities in Reissner's Theory for the Bending of Elastic Plates, J. Appl. Mech. 53 (1986) pp.220-222.

12. P.G. Ciarlet P. Destuynder, A Justification of the Two Dimensional Linear Plate Model, J. Mécanique 18 (1979), 315-344.

13. J.L. Chaboche, Time Independent Constitutive Theorie for Cycle Plassticity Int. J. of Plasticity 2 (1986), 149-188.

14. M. Costabel E. Stephan, Curvature Terms in the Asymptotic Expansions for Solutions of Boundary Integral Equations on Curved Polygons of Integral Equations 5(1983), 353-371.

15. M. Dauge, Elliptic Boundary Value Problems on Corner Domains. Lecture Notes in Math. 1341, Springer, New York 1988.

16. P. Grisrard Elliptic Problems in Nonsmooth Domains, Pitman, Boston 1985.

17. W. Han. Error Estimations of the Idealization of Mathematical Formulations of Problems Leading to Elliptic Partial Differential Equations, Ph.D. Thesis (1990), University of Maryland, College Park, MD 20742, USA.

18. K. Jerina, I. Babuška Mathematical Modeling of Physical Systems with Considerations of Uncertainities, Washington University, Mech. Engtg. Dept. St.Louis, 1990.

19. D. Morgenstern, Herleitunng der Plattenthearie aus der Dreidimensionales Elastizitätstheorie, Arch Rat Mech. Anal. 4 (1959), 145-152.

20. A.K. Noor, W.S. Burton, Assessment of Shear Deformation Theories for Multilayered Composite Plates, Appl. Mech. Rev. 42 (1989), pp.1-12.

21. A. Papadakis, Computational aspects of Determination of Stress Intensity Factors for Two-dimensional Elasticity, Ph.D. Thesis, (1988), University of Maryland, College Park, MD 20742, USA.

22. T. Petersdorff, Randwertprobleme der Elastizitätstheorie für Polyeder - singularitäten und Approximation mit Randenelement methodem, Ph.D. Thesis (1989), T.H. Darmstadt FRG.

23. E. Reissner. Reflections on Theory of Elastic Plates, Appl. Math. Rev. 38 (1985), pp.1453-1464.

24. C. Schwab. The Dimensional Reduction Method, Ph.D. Thesis (1989), University of Maryland, College Park, MD 20742, USA.

25. B.A. Szabo, I. Babuška, Computation of the Amplitude of Stress Singular Terms for Cracks and reendrant Corners, Fracture Mechanics XIX Symp. ASTM STP 1969, T.A. Cruse ed., Am Soc. Test.and Mat.,Philadelphia, 1987, pp. 101-126.

26. M. Vogelius, I. Babuška, On a Dimensional Reduction Method, Math. of Comp. 37(1981), pp. 31-45, 47-68.

27. D.H.Y. Yen, M. Zhow, On the Singularity of Corner Points of Solutions of Plate Bending Problem, Preprint, Department of Mathematics, Michigan State University, East Lansing, 1989.

28. M. Zyczkowski, Combined Loadings in the Theory of Plasticity, PWN-Polish Scientific Publ. Warsazava 1981.

A Posteriori Error Analysis In Finite Elements: The Element Residual Method For Symmetrizable Problems With Applications to Compressible Euler and Navier-Stokes Equations

J. T. Oden[*] L. Demkowicz[*] W. Rachowicz[†] T. A. Westermann[†]

Abstract

An extension of the element residual method for *a posteriori* error estimation to symmetrizable problems is presented. Applications include compressible Euler and Navier-Stokes equations.

1 Introduction

The interest in *a posteriori* error estimation in finite element methods began with a series of papers by Babuška and his collaborators in the late seventies (see [1] for sample results related to this presentation, or [10] for a more complete list of references) and resulted in the first conference devoted to adaptivity and reliability of finite element computations in Lisbon, 1984. By that time a few generalizations of existing techniques were investigated including the so-called *Element Residual Method* (ERM) proposed independently by Bank and Weiser in [3] and Oden and Demkowicz in [6, 11]. Recently, in [12], the method was extended to arbitrary, combined *h-p* approximations and compared (favorably) with other *a posteriori* error estimation techniques in the context of a model elliptic problem.

As current interest in adaptive methods extends to problems in fluid dynamics, the natural question arises as to whether the method can be extended to handle more general

[*] The Texas Institute for Computational Mechanics, The University of Texas at Austin
[†] The Computational Mechanics Company, Inc.

problems, including those of compressible gas dynamics. In particular, many basic issues arise in extending these methods to nonlinear, unsymmetrical operators, such as, for example, the choice of a norm to be used in place of the energy norm that arises naturally in the context of elliptic problems ([13]).

In this paper, we propose a generalization of the ERM to symmetrizable problems which includes such problems of interest as the time-step-dependent boundary value problems resulting from the time discretization of the Euler or Navier-Stokes equations. The natural norm is then identified as the *linearized entropy* corresponding to a particular solution vector (steady state solution for steady state problems).

The idea of symmetrization can be traced back to the works of Friedrichs (see, e.g., [7]). Recently, the relationship of symmetrization to the notion of entropy functions (see [5] and references therein) for the Euler equations (see Hughes, et al. [6]) and for the Navier-Stokes equations was established.

The numerical examples presented in this paper are based on the time discretization schemes for Euler and Navier-Stokes equations presented in [4, 5].

2 Element Residual Method

Given a domain $\Omega \subset \mathbb{R}^N$ (we assume $N = 2$ for notational simplicity) we consider a general variational boundary value problem in the form

$$\left\{ \begin{array}{l} \text{Find } \boldsymbol{U} \in \boldsymbol{X} \text{ such that} \\ B(\boldsymbol{U}, \boldsymbol{W}) = L(\boldsymbol{W}) \text{ for every } \boldsymbol{W} \in \boldsymbol{X} \end{array} \right. \tag{2.1}$$

where

$$\begin{aligned} \boldsymbol{X} = \boldsymbol{H}^1(\Omega) &= \underbrace{H^1(\Omega) \times \ldots \times H^1(\Omega)}_{n \text{ times}} \\ B(\boldsymbol{U}, \boldsymbol{W}) &= \sum_{i,j=1}^{n} B^{ij}(U_i, W_j) \qquad (2.2) \\ L(\boldsymbol{W}) &= \sum_{j=1}^{n} L^j(W_j) \end{aligned}$$

with the bilinear forms B^{ij} and linear forms L^j defined as (omitting superscripts for notational convenience)

$$\begin{aligned} B(u,w) &= \int_\Omega \left\{ \sum_{k,\ell=1}^{2} a_{kl} \frac{\partial u}{\partial x_\ell} \frac{\partial w}{\partial x_k} + \sum_{k=1}^{2} b_k \frac{\partial u}{\partial x_k} w + \sum_{\ell=1}^{2} d_\ell u \frac{\partial w}{\partial x_\ell} + cuv \right\} dx \\ &+ \int_{\partial\Omega} \left\{ b_s \frac{\partial u}{\partial s} w + d_s u \frac{\partial w}{\partial s} + c_s uw \right\} ds \end{aligned} \tag{2.3}$$

$$\begin{aligned} L(w) &= \int_\Omega \left\{ fw + \sum_{\ell=1}^{2} g_\ell \frac{\partial w}{\partial x_\ell} \right\} dx \\ &+ \int_{\partial\Omega} f_s w \, ds \end{aligned} \tag{2.4}$$

For each pair of indices $i,j = 1,\ldots,n$, a_{kl}, b_k, d_ℓ, c, f, g_ℓ are functions specified in Ω and b_s, d_s, c_s, f_s are functions specified on the boundary $\partial\Omega$. The normal and tangential derivatives on the boundary are defined as

$$\begin{aligned} \frac{\partial u}{\partial n} &= \frac{\partial u}{\partial x_1} n_1 + \frac{\partial u}{\partial x_2} n_2 \\ \frac{\partial u}{\partial s} &= \frac{\partial u}{\partial x_1}(-n_2) + \frac{\partial u}{\partial x_2} n_1 \end{aligned} \tag{2.5}$$

where (n_1, n_2) are components of the outward normal unit vector $\boldsymbol{n}$.

Systems cf type (2.1) include not only classical elliptic equations of second-order but also arise naturally as "one time step problems" from different time discretization schemes applied to parabolic or hyperbolic equations. The boundary integrals in (2.3) permit the implementation of different boundary conditions (including Dirichlet boundary conditions via the penalty method).

Replacing $\boldsymbol{X}$ in (2.1) with a finite dimensional subspace $\boldsymbol{X}_{h,p}$ of $\boldsymbol{X}$ we arrive at the approximate problem

$$\left\{ \begin{array}{l} \text{Fir ' } \boldsymbol{U}_{h,p} \in \boldsymbol{X}_{h,p} \text{ such that} \\ \\ B(\boldsymbol{U}_{h,p}, \boldsymbol{W}) = L(\boldsymbol{W}) \qquad \forall\, \boldsymbol{W} \in \boldsymbol{X}_{h,p} \end{array} \right. \tag{2.6}$$

Indices h and p refer here to the use of an arbitrary h-p adaptive finite element (FE) meshes, with locally varying mesh size h and spectral order of approximation p extensively studied in [12].

It is our goal to propose and investigate here a general method for estimating the *relative residual error* corresponding to (2.6). More precisely, considering the enriched space $\boldsymbol{W}_{h,p+1}$ corresponding to the same mesh but with local order of approximation uniformly increased by one, we define the relative residual error as

$$\sup_{\boldsymbol{W} \in \boldsymbol{X}_{h,p+1}} \frac{|B(\boldsymbol{U}_{h,p}, \boldsymbol{W}) - L(\boldsymbol{W})|}{\|\boldsymbol{W}\|} \tag{2.7}$$

The choice of norm $\|\boldsymbol{W}\|$ is unfortunately not unique. Two important special cases are, however, of interest: the *symmetric case*, when B is symmetric and positive definite, and the *symmetrizable case* when B can be made symmetric by an appropriate change of variables.

Symmetric Case

When the bilinear form B is symmetric and positive definite and the energy norm

$$\|\boldsymbol{W}\|_E = B(\boldsymbol{W}, \boldsymbol{W}) \tag{2.8}$$

is selected in (2.7) the residual error is *equal* to the relative error between $\boldsymbol{U}_{h,p}$ and $\boldsymbol{U}_{h,p+1}$, the FE solution corresponding to the enriched space and (see [11]) measured in the energy norm.

$$\sup_{\boldsymbol{W} \in \boldsymbol{X}_{h,p+1}} \frac{|B(\boldsymbol{U}_{h,p}, \boldsymbol{W}) - L(\boldsymbol{W})|}{\|\boldsymbol{W}\|_E} = \|\boldsymbol{U}_{h,p} - \boldsymbol{U}_{h,p+1}\|_E \tag{2.9}$$

The principal idea behind the proposed error estimate is to interpret (2.9) as a variational formulation of an elliptic problem, transform the bilinear form $\boldsymbol{B}$ into the typical form for elliptic problems, and finally apply the element residual method presented in [12].

Formally, we proceed as follows:

Step 1: Transform formulas (2.3) and (2.4) into the typical form for elliptic equations.

$$\begin{aligned} B(u,w) &= \int_\Omega \left\{ \sum_{k,l=1}^{2} a_{kl} \frac{\partial u}{\partial x_\ell} \frac{\partial w}{\partial x_k} \right. \\ &+ \left. \sum_{k=1}^{2} (b_k - d_k) \frac{\partial u}{\partial x_k} w + \left(c - \sum_{l=1}^{2} \frac{\partial d_l}{\partial x_l} \right) uw \right\} dx \\ &+ \int_{\partial\Omega} \left\{ b_s \frac{\partial u}{\partial s} w + d_s u \frac{\partial w}{\partial s} + \left(c_s + \sum_{\ell=1}^{2} d_l n_l \right) uw \right\} ds \\ L(w) &= \int_\Omega \left(f - \sum_{\ell=1}^{2} \frac{\partial g_\ell}{\partial x_\ell} \right) w dx + \int_{\partial\Omega} \left(f_s + \sum_{\ell=1}^{2} g_\ell n_\ell \right) w ds \end{aligned} \tag{2.10}$$

Step 2: Apply the element residual method to the modified bilinear and linear forms resulting in the estimate

$$\|\boldsymbol{U}_{h,p} - \boldsymbol{U}_{h,p+1}\|_E \leq \left(\sum_K \|\varphi_K\|^2_{E,K}\right)^{\frac{1}{2}} \tag{2.11}$$

where the *error indicator function* φ_K is the solution to the local problem

$$\left\{\begin{array}{l}
\text{Find } \varphi_K \in \boldsymbol{X}^0_{h,p+1}(K) \text{ such that} \\[1em]
B_K(\varphi_K, \boldsymbol{W}) = \\
= \displaystyle\sum_{j=1}^{n} \Big\{ \int_K \Big\{ f^j - \sum_{i=1}^{2} \frac{\partial g^j_l}{\partial x_l} \\
\qquad - \displaystyle\sum_{i=1}^{n} \Big[- \sum_{k,l=1}^{2} \frac{\partial}{\partial x_k}\left(a^{ij}_{kl} \frac{\partial u^i}{\partial x_l}\right) - \sum_{k=1}^{2} \left(b^{ij}_k - d^{ij}_k \right) \frac{\partial u^i}{\partial x_k} \\
\qquad\qquad - \displaystyle\left(c^{ij} - \sum_{l=1}^{2} \frac{\partial d^{ij}_l}{\partial x_l} \right) u^i \Big] \Big\} w^j dx \\
\qquad + \displaystyle\int_{\partial K \backslash \partial \Omega} \left\{ \left[\!\left[\sum_{i=1}^{2} \sum_{k,l=1}^{2} a^{ij}_{kl} \frac{\partial u^i}{\partial x_l} n_k \right]\!\right] - \sum_{i=1}^{2} \sum_{k,l=1}^{2} a^{ij}_{kl} \frac{\partial u^i}{\partial x_l} n_k \right\} w^j ks \\
\qquad \displaystyle\int_{\partial K \cup \partial \Omega} \Big\{ f^j_s + \sum_{l=1}^{2} g^j_l n_l \\
\qquad - \displaystyle\sum_{i=1}^{n} \left[b^{ij}_s \frac{\partial u^i}{\partial s} w^j + d^{ij}_s u^i \frac{\partial w^j}{\partial s} + c^{ij}_s u^i w^j \right] \Big\} ds \Big\} \\[1em]
\text{for every } \boldsymbol{W} \in \boldsymbol{X}^0_{h,p+1}(K)
\end{array}\right. \tag{2.12}$$

Here $\boldsymbol{X}^0_{h,p+1}(K)$ is the kernel of the *h-p* interpolation operator defined on the element enriched space $\boldsymbol{X}_{h,p+1}(K)$ or the so-called space of element bubble functions (see [12] for details) and the element bilinear form B_K is defined as the element contribution to (2.10). Finally, the symbol $[\![\ ]\!]$ denotes the average flux defined along the interelement boundary and evaluated using both the element and the neighboring elements values of derivatives and coefficients a^{ij}_{kl} (if they are discontinuous). The element energy in (2.11) is defined using the element bilinear form B_K.

Step 3: Integrating by parts, transform the element bilinear form and the right-hand side of the local problem into the form consistent with the initial formulas for B and L.

We arrive at the following formulas

$$\begin{aligned} B_K(\varphi, \boldsymbol{W}) &= \sum_{i,j=1}^{n} B_K^{ij}(\varphi^i, w^j) \\ L_K(\boldsymbol{W}) &= \sum_{j=1}^{n} L_K^j(w^j) \end{aligned} \tag{2.13}$$

where

$$\begin{aligned} B_K(\varphi, w) &= \int_K \left\{ \sum_{k,l=1}^{2} a_{kl} \frac{\partial \varphi}{\partial x_\ell} \frac{\partial w}{\partial x_k} + \sum_{k=1}^{2} b_k \frac{\partial \varphi}{\partial x_k} w + \sum_{\ell=1}^{2} d_\ell \varphi \frac{\partial w}{\partial x_\ell} + c\varphi w \right\} dx \\ &+ \int_{\partial K \setminus \partial\Omega} - \left(\sum_{\ell=1}^{2} g_\ell n_\ell \right) \varphi w \; ds \\ &+ \int_{\partial K \cap \partial\Omega} \left\{ b_s \frac{\partial \varphi}{\partial s} w + d_s \varphi \frac{\partial w}{\partial s} + c_s \varphi w \right\} ds \\ L_K(w) &= \int_K \left(fw + \sum_{\ell=1}^{2} g_\ell \frac{\partial w}{\partial x_\ell} \right) dx \\ &- \int_{\partial K \setminus \partial\Omega} \left\{ \sum_{\ell=1}^{2} g_\ell n_\ell \right) w \; ds \\ &+ \int_{\partial K \cap \partial\Omega} f_s w ds \end{aligned} \tag{2.14}$$

The final form of the local problem is derived as follows:

$$\left\{ \begin{aligned} & \text{Find } \varphi_K \in \boldsymbol{X}^0_{h,p+1}(K) \text{ such that} \\ & B_K(\varphi_K, \boldsymbol{W}) = L_K(\boldsymbol{W}) - B_K(\boldsymbol{U}_{h,p}, \boldsymbol{W}) \\ & \qquad + \sum_{i,j=1}^{n} \int_{\partial K \setminus \partial\Omega} \left[\!\left[\sum_{k,l=1}^{2} a_{kl}^{ij} \frac{\partial u^i}{\partial x_\ell} \right]\!\right] w^j ds \end{aligned} \right. \tag{2.15}$$

Nonsymmetric and Symmetrizable Problems

Formally, formula (2.11) can be used for nonsymmetric problems as well, as long as the local element bilinear forms B_K are positive semidefinite, i.e.,

$$B_K(\varphi_K, \varphi_K) \geq 0 \tag{2.16}$$

This happens if the symmetric contributions to B_K dominate the unsymmetric ones (resulting usually from the first-order terms). The global bilinear form B is then automatically semipositive and, with the correct boundary conditions, it is positive definite. This guarantees the well-posedness of the problem.

Another interesting case is when the bilinear form is nonsymmetric but it is *symmetrizable* in the sense that a matrix-valued function $\boldsymbol{A}_0(\boldsymbol{x})$ exists (the so-called symmetrizer) such that a new bilinear form $\tilde{B}$ defined as

$$\tilde{B}(\boldsymbol{U},\boldsymbol{W}) = B(\boldsymbol{U},\boldsymbol{A}_0\boldsymbol{W}) \tag{2.17}$$

is symmetric.

If, in addition, the symmetrized bilinear form $\tilde{B}$ is positive definite, then the error estimation technique can be extended to this case as well.

Remark: Note that in this case the original bilinear form satisfies the inf-sup stability condition

$$\inf_{\|\boldsymbol{U}\|=1} \sup_{\boldsymbol{W}\neq 0} \frac{|B(\boldsymbol{U},\boldsymbol{W})|}{\|\boldsymbol{W}\|} > \alpha > 0 \tag{2.18}$$

(It is enough to take $\boldsymbol{W} = \boldsymbol{A}_0\boldsymbol{U}$). ■

Introduction of the symmetrizer *does not effect* the construction and solution of the local problems. It only helps identify the norm for the space $\boldsymbol{X}_{h,p+1}$ in (2.7) and affects the evaluation of the error estimate. Using the same definition of element bilinear and linear forms B_K, L_K, we proceed as follows:

Step 1: Use the orthogonality of the residual to the $\boldsymbol{X}_{h,p}$ space,

$$B(\boldsymbol{U}_{h,p},\boldsymbol{W}) - L(\boldsymbol{W}) = B(\boldsymbol{U}_{h,p},\boldsymbol{\psi}) - L(\boldsymbol{\psi}) \tag{2.19}$$

where

$$\boldsymbol{\psi} = \boldsymbol{W} - \Pi_{h,p}\boldsymbol{W} \tag{2.20}$$

where $\Pi_{h,p}$ denotes the *h-p* interpolation operator (see [12]).

Step 2: Decompose the bilinear and linear forms according to formulas (2.10) introducing the average flux interelement boundary terms

$$\begin{aligned} &B(\boldsymbol{U}_{h,p},\boldsymbol{W}) - L(\boldsymbol{W}) \\ &= \sum_K B_K(\boldsymbol{U}_{h,p},\boldsymbol{\psi}) - L_K(\boldsymbol{\psi}) + \sum_{i,j=1}^{n} \int_{\partial K\setminus\partial\Omega} \left[\!\left[\sum_{k,l=1}^{2} a_{kl}^{ij} \frac{\partial u^i}{\partial x_l} n_k \right]\!\right] \psi^j \, ds \end{aligned} \tag{2.21}$$

Step 3: Introduce the solutions to the local problems

$$B(U_{h,p}, W) - L(W) = \sum_K B_K(\varphi_K, \psi_K) \tag{2.22}$$

where ψ_K is the restriction of ψ to element K.

Step 4: Introduce the symmetrizer and use the Cauchy-Schwartz inequality for the symmetrized form to estimate the error

$$\begin{aligned} B(U_{h,p}, W) - L(W) &= \sum_K B_K(\varphi_K, A_0 A_0^{-1}\psi_K) \\ &= \sum_K \tilde{B}_K(\varphi_K, A_0^{-1}\psi_K) \leq \sum_K \tilde{B}(\varphi_K, \varphi_K)^{\frac{1}{2}} \tilde{B}_K(A_0^{-1}\psi_K, A_0^{-1}\psi_K)^{\frac{1}{2}} \\ &\leq C\left[\sum_K B_K(\varphi_K, A_0\varphi_K)\right]^{\frac{1}{2}} B(A_0^{-1}W, W)^{\frac{1}{2}} \end{aligned} \tag{2.23}$$

Here $C = \max_K C_K$ where for every element K, C_K is identified as the norm of $(I - \Pi_{h,p})$ operator with respect to the element energy norm defined as

$$\|W\|^2_{E,K} = B_K(A_0^{-1}W, W) \tag{2.24}$$

(see [12] for a detailed discussion of C). For undistorted meshes C is close to one (independent of the order of approximation!) and in practical calculations is neglected.

Identifying the global energy norm for W in (2.7) as the sum of (2.24) we arrive at the final estimate of the form

$$\sup_{W \in X_{h,p+1}} \frac{|B(U_{h,p}, W) - L(W)|}{\|W\|} \leq \left[\sum_K B_K(\varphi_K, A_0\varphi_K)\right]^{\frac{1}{2}} \tag{2.25}$$

3 Examples of Symmetrizable Problems

Linear Diffusion Problem in "Momentum Components"

We start with a model elliptic system of linear elasticity equations

$$\begin{aligned} -\frac{\partial}{\partial x_1}\left[(2\mu+\lambda)\frac{\partial u_1}{\partial x_1} + \lambda\frac{\partial u_2}{\partial x_2}\right] - \frac{\partial}{\partial x_2}\left[\mu\left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}\right)\right] &= f_1 \\ -\frac{\partial}{\partial x_1}\left[\mu\left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}\right)\right] - \frac{\partial}{\partial x_2}\left[\lambda\frac{\partial u_1}{\partial x_1} + (2\mu+\lambda)\frac{\partial u_2}{\partial x_2}\right] &= f_2 \end{aligned} \tag{3.1}$$

with (for simplicity only) Dirichlet boundary conditions

$$u_1 = \hat{u}_1 \ , \ u_2 = \hat{u}_2 \quad \text{on } \partial\Omega \tag{3.2}$$

Above, $\boldsymbol{f} = (f_1, f_2)$ is the body force vector and μ and λ are the shear and bulk viscosities. Multiplying (3.1) by a test function $\boldsymbol{v} = (v_1, v_2)$ and using the penalty method approach for (3.2) we arrive at the classical variational formulation in the form

$$\left\{ \begin{array}{ll} \text{Find } \boldsymbol{u} \in \boldsymbol{X} & \text{such that} \\ & \\ B(\boldsymbol{u}, \boldsymbol{v}) = L(\boldsymbol{v}) & \forall\, \boldsymbol{v} \in \boldsymbol{X} \end{array} \right. \tag{3.3}$$

where $\boldsymbol{X} = \boldsymbol{H}^1(\Omega)$ and

$$\begin{aligned} B(\boldsymbol{u}, \boldsymbol{v}) &= \sum_{i,j=1}^{2} B^{ij}(u_i, v_j) \\ L(\boldsymbol{v}) &= \sum_{j=1}^{2} L^j(v_j) \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} B^{11}(u_1, v_1) &= \int_\Omega \left\{ (2\mu + \lambda) \frac{\partial u_1}{\partial x_1} \frac{\partial v_1}{\partial x_1} + \mu \frac{\partial u_1}{\partial x_2} \frac{\partial v_1}{\partial x_2} \right\} ds \\ &+ \int_{\partial\Omega} \frac{1}{\varepsilon} u_1 v_1 \, ds \\ B^{12}(u_2, v_1) &= \int_\Omega \left\{ \mu \frac{\partial u_2}{\partial x_1} \frac{\partial v_1}{\partial x_2} + \lambda \frac{\partial u_2}{\partial x_2} \frac{\partial v_1}{\partial x_1} \right\} dx \\ B^{21}(u_1, v_1) &= \int_\Omega \left\{ \mu \frac{\partial u_1}{\partial x_2} \frac{\partial v_2}{\partial x_1} + \lambda \frac{\partial u_1}{\partial x_1} \frac{\partial v_2}{\partial x_2} \right\} dx \\ B^{22}(u_2, v_2) &= \int_\Omega \left\{ \mu \frac{\partial u_2}{\partial x_1} \frac{\partial v_2}{\partial x_1} + (2\mu + \lambda) \frac{\partial u_2}{\partial x_2} \frac{\partial v_2}{\partial x_2} \right\} dx \\ &+ \int_{\partial\Omega} \frac{1}{\varepsilon} u_2 v_2 \, ds \\ L_1(v_1) &= \int_\Omega f_1 v_1 dx + \int_{\partial\Omega} \frac{1}{\varepsilon} \hat{u}_1 v_1 ds \\ L_2(v_2) &= \int_\Omega f_2 v_2 dx + \int_{\partial\Omega} \frac{1}{\varepsilon} \hat{u}_2 v_2 ds \end{aligned} \tag{3.5}$$

Obviously, the bilinear form is symmetric and positive definite.

Strange as it looks, we may be interested in solving (3.1) in terms of new variables

$$m_i = \rho u_i \tag{3.6}$$

where $\rho = \rho(x_1, x_2)$ is a specified, continuous function. A situation like this happens when the compressible Navier-Stokes equations in conservative variables are discretized in time by means of the operator-splitting approach.

Substitution of $u_i = 1/\rho m_i$ into (3.1) and consequently into the variational formulation results in an *unsymmetric* formulation due to the presence of first-order terms as

$$\frac{\partial u_i}{\partial x_j} = \frac{1}{\rho}\frac{\partial m_i}{\partial x_j} - \frac{1}{\rho^2}\frac{\partial \rho}{\partial x_j} m_i \tag{3.7}$$

Obviously the symmetry of the problem can be recovered by making an identical substitution for the test function, i.e., introducing the symmetrizer of the form

$$A_0 = \begin{pmatrix} \frac{1}{\rho} & 0 \\ 0 & \frac{1}{\rho} \end{pmatrix} \tag{3.8}$$

Taylor-Galerkin Method for Euler Equations

Consider the compressible gas dynamics equations in the form

$$\boldsymbol{U}_{,t} + \frac{\partial}{\partial x_i}\boldsymbol{F}^i(\boldsymbol{U}) = 0 \tag{3.9}$$

where $\boldsymbol{U}$ is the vector of conservative variables (density, momentum components, energy) and $\boldsymbol{F}^i$ are the Euler fluxes (algebraic functions of $\boldsymbol{U}$). Starting with the second-order finite difference formula in time

$$\boldsymbol{U}(t+\Delta t) - \frac{\Delta t^2}{2}\boldsymbol{U}_{,tt}(t+\Delta t) = \boldsymbol{U}(t) + \Delta t \boldsymbol{U}_{,t}(t) + 0(\Delta t^3) \tag{3.10}$$

we use the original equations (3.9) to represent the time derivatives in terms of spatial derivatives and arrive at a simple time step problem of the form

$$\boldsymbol{U}^{n+1} - \frac{\Delta t^2}{2}\sum_{k,l=1}^{2}\frac{\partial}{\partial x_k}\left(A_k A_\ell \frac{\partial \boldsymbol{U}^{n+1}}{\partial x_l}\right) = \boldsymbol{U}^n - \Delta t \frac{\partial}{\partial x_l}\boldsymbol{F}^l(\boldsymbol{U}) \tag{3.11}$$

where the Jacobian matrices $\boldsymbol{A}_k = \boldsymbol{F}^k_{\boldsymbol{U}}$ are evaluated at $\boldsymbol{U}^n$.

Multiplying (3.11) by a vector–valued test function W, integrating over Ω and integrating by parts we arrive at the variational formulation of the form (2.1) with the bilinear and linear forms defined as follows

$$\begin{aligned} B\left(U^{n+1}, W\right) &= \int_\Omega \left\{ W^T U^{n+1} + \frac{\Delta t^2}{2} \sum_{k,l=1}^{2} \left(\frac{\partial W}{\partial x_k}\right)^T A_k A_l \frac{\partial U^{n+1}}{\partial x_\ell} \right\} dx \\ &+ \text{ boundary terms} \\ L(W) &= \int_\Omega \left\{ W^T U^n + \Delta t \sum_{k=1}^{2} \left(\frac{\partial W}{\partial x_k}\right)^T F^k(U^n) \right\} dx \\ &- \int_{\partial\Omega} \Delta t W^T \left(\sum_{k=1}^{2} F^k(U^n) n_k \right) ds \end{aligned} \tag{3.12}$$

The form of boundary terms present in the formula for the bilinear form depends on boundary conditions (see [5] for a detailed discussion).

The formulation is *nonsymmetric.* However, it is known (see [7, 8]) that there exists a symmetrizer $A_0 = A_0(U)$ (Hessian of the entropy function for Euler equations) such that

$$\begin{aligned} &1. \quad A_0 = A_0^T > 0 \\ &2. \quad (A_0 A_i)^T = A_0 A_i \geq 0 \end{aligned} \tag{3.13}$$

Based on (3.13), one can easily verify that (with a proper treatment of boundary conditions) the bilinear form

$$\tilde{B}(U, W) = B(U, A_0 W) \tag{3.14}$$

is symmetric, provided the derivatives of the symmetrizer A_0 are negligible, i.e.,

$$\frac{\partial}{\partial x_i} A_0 \approx 0 \tag{3.15}$$

The explicit formula for the symmetrizer is given in Fig. 1.

Operator–Splitting Method for Navier-Stokes Equations

1. *The Convection Operator*

We define

$$\left(E(t) U_0\right)(x) \overset{\text{def}}{=} U(x, t) \tag{3.16}$$

$$A_0 = \begin{bmatrix} \frac{1}{\rho}[\psi^2 + \gamma] & -\psi\frac{u}{\iota} & -\psi\frac{v}{\iota} & \frac{1}{\iota}(\psi - 1) \\ & \frac{1}{\iota}\left(1 + \rho\frac{u^2}{2}\right) & \rho\frac{uv}{\iota^2} & -\rho\frac{u}{\iota^2} \\ \text{Sym.} & & \frac{1}{\iota}\left(1 + \rho\frac{v^2}{\iota}\right) & -\frac{\rho v}{\iota^2} \\ & & & \frac{\rho}{\iota^2} \end{bmatrix}$$

where $\psi = \rho\frac{u^2 + v^2}{2\iota}$ with ρ the density, u and v the velocity components, and ι the internal energy (per unit volume).

Figure 1: The symmetrizer.

where $U(x,t)$ is the solution to the system of Euler equations (transport step)

$$U_{,t} + \sum_{i=1}^{2} F^i(U)_{,i} = 0 \tag{3.17}$$

with the initial condition

$$U(x,0) = U_0(x) \tag{3.18}$$

anc. appropriate boundary conditions.

2. *The Diffusion Operator*

$$\left(H(t)U_0\right)(x) \overset{\text{def}}{=} U(x,t) \tag{3.19}$$

where $U(x,t)$ is the solution to the system of equations (viscous step) in the form

$$U_{,t} = \sum_{i=1}^{2} \left(\sum_{j=1}^{2} K^{ij}(U)U_{,j} \right)_{,i} \tag{3.20}$$

with initial condition (3.18) and appropriate boundary conditions.

Explicit formulas for the viscous fluxes $F_i^V = \sum_{j=1}^2 K_{ij}U_{,j}$ and the viscous matrices K_{ij} can be found in [4].

Two compositions of the operators H and E may be considered, a two-step splitting of the form

$$G(t) = H(t)E(t) \tag{3.21}$$

and a three-step Strang procedure of the form

$$S(t) = H\left(\frac{t}{2}\right) E(t) H\left(\frac{t}{2}\right) \tag{3.22}$$

It can be shown that the first procedure is of first-order while the second is of second-order in time, i.e.,

$$\begin{aligned} \left|G(t)U_0 \;-\; U(x,t)\right| &\le c(x)t^2 \\ \left|S(t)U_0 \;-\; U(x,t)\right| &\le c(x)t^3 \end{aligned} \tag{3.23}$$

where $U(x,t)$ is a solution to the full Navier-Stokes equations with initial condition (3.18) and $c(x)$ is an unknown function of x. Practically speaking, the solution of a single time step problem breaks into two fractional steps for the first-order splitting and three for the second-order splitting methods. The transport step reduces to the Euler equations and is solved using the Taylor-Galerkin method. The form of the differential equation's (3.20) defining the viscous step leads to two simple observations:

- The density ρ remains unchanged in the viscous step, i.e.,

$$\rho^{n+1} = \rho^{n} \tag{3.24}$$

- The remaining three equations can be decoupled. We first solve a system of two equations for the momentum components m_i

$$m_{i,t} = \sum_{j=1}^{2} (\tau_{ij})_{,i} \tag{3.25}$$

  and then a single equation for the total energy e

$$e_{,t} = \sum_{j=1}^{2} \tau_{ij} u_j + q_i \tag{3.26}$$

To affect the decoupling, the boundary conditions for (3.25) must be formulated in such a way that they do not contain energy terms.

As a starting point to solve both (3.25) and (3.26), we accept a first-order finite difference formula of the form

$$U(t+\Delta t) - \beta \Delta t U_{,t}(t+\Delta t) = U(t) + (1-\beta)\Delta t U_{,t}(t) + 0(\Delta t^2) \tag{3.27}$$

Note that for $\beta = \frac{1}{2}$, (3.27) is of second-order and reduces to a Crank-Nicholson scheme.

Replacing the time derivatives with spatial derivatives in (3.25), (3.26) is approximated with a system of two equations of the form

$$m_j^{n+1} - \beta \Delta t \sum_{i=1}^{2} \tau_{ij,i}^{n+1} = m_j^n + (1-\beta)\Delta t \sum_{i=1}^{1} \tau_{ij,i}^{n} \tag{3.28}$$

Equations (3.28), if rewritten in terms of the velocity components, reduce to a system of two symmetric, elliptic equations. Unfortunately, in order to comply with the conservative form of the equations, (3.28) *must be solved* in momentum components.

As a next step, equations (3.28) are linearized by evaluating the viscosities μ and λ (*not* the whole viscous matrices $\boldsymbol{K}^{ij}$) as functions of $\boldsymbol{U}^n$ rather than $\boldsymbol{U}^{n+1}$. Once the system (3.28) is solved, one proceeds with the energy equation. Combining (3.27) with (3.26) gives a single equation of the form

$$\begin{aligned} e^{n+1} - \beta \Delta t \sum_{i=1}^{2} \left(\sum_{j=1}^{2} \tau_{ij}^{n+1} u_j^{n+1} + \kappa \theta_{,i}^{n+1} \right)_{,i} \\ = e^{n} + (1-\beta)\Delta t \sum_{i=1}^{2} \left(\sum_{j=1}^{2} \tau_{ij}^{n} u_j^{n} + \kappa \theta_{,i}^{n} \right)_{,i} \end{aligned} \tag{3.29}$$

which, when rewritten in terms of temperature, (3.29) takes the form of a single, elliptic equation. In their variational formulations neither (3.28) nor (3.29) result in symmetric problems but in both cases the bilinear forms may be symmetrized using the symmetrizer (3.8) for momentum equations and $A_0 = 1/\rho$ for the energy equation.

Let us notice finally that had we not decoupled the momentum equations from the energy one, the same symmetrizer A_0 as for the Euler equations could be used (comp. [8]).

4 Numerical Examples

In this section, three example problems using these techniques of error estimation are presented. The results take the form of plots of the error estimates and effectivity indices as well as global effectivity indices and standard deviations. These quantities are defined as follows:

$$\gamma_K = \frac{\theta_K}{|||e|||_K} \tag{4.1}$$

where γ_K is the effectivity index for element K, θ_K is the estimated error and $|||e|||_K$ is the actual element error in the coarse mesh approximation (comparing the coarse mesh approximation with either the analytic solution or the approximate solution on a mesh of uniformly increased polynomial order). Additionally, we introduce a discrete measure (weight) ω_K defined according to

$$\omega_K = \frac{|||e|||_K^2}{|||e|||^2} \tag{4.2}$$

the global effectivity index becomes:

$$\gamma^2 = \frac{\theta^2}{|||e|||^2} = \frac{\left(\sum_K \theta_K^2\right)}{|||e|||^2} = \sum_K \gamma_K^2 \omega_K \tag{4.3}$$

Now classical statistics suggest a standard deviation σ (with respect to the measure) as a method to quantify the ability of the estimates to predict an appropriate distribution of error. The standard deviation is defined as:

$$\sigma^2 = \sum_K \left(\gamma_K^2 - \gamma^2\right)^2 \omega_K \tag{4.4}$$

In order to eliminate any global constants that may be missing from our estimates, we normalize the element effectivity indices by dividing them by the global effectivity index:

$$\overline{\gamma}_K = \frac{\theta_K}{|||e|||_K} \cdot \gamma^{-1} \tag{4.5}$$

which results in a standard deviation defined according to:

$$\overline{\sigma}^2 = \sum_K \left(\overline{\gamma}_K^2 - 1\right)^2 \omega_K \tag{4.6}$$

Example 1: Linear Elasticity Step in "Momentum Components"

The problem discussed in Section 3 was solved on the L-shaped domain shown in Fig. 2 with homogenous boundary conditions. The fluid viscosities were chosen $\lambda = 4.0$, $\mu = 1.0$ and f_1 and f_2 consistent with the analytic solution:

$$\begin{aligned} u_1 &= \frac{r}{r^2 + 0.01} \sin\theta(1 - x^2)(y - y^3) \\ u_2 &= \frac{-r}{r^2 + 0.01} \cos\theta(x - x^3)(1 - y^2) \end{aligned} \tag{4.7}$$

(where r and θ are defined in Fig. B.1). Additionally, we have chosen:

$$\rho = (1.1 - x_1)(1.1 - x_2) \tag{4.8}$$

Figure 3 shows plots of the estimated errors and normalized effectivity indices corresponding to a mesh of 48 quadratic elements. For this problem the global effectivity index was 1.276 and the standard deviation 0.537. Notice that the range of effectivity indices shown is 0.0 to 2.0. (Elements with an effectivity index greater than 2.0 are assigned the darkest shade.)

Example 2: Flow Over a Blunt Body Problem for the Euler Equations

We used the Taylor-Galerkin method described in Section 3 to solve a flow over a blunt body problem with Mach number $Ma = 6$. Figure 4 shows the density contours of a steady-state solution obtained on a uniform mesh of 16 × 16 linear elements. Figures 5a and 5b present distributions of error indicators θ_K and normalized effectivity indices $\overline{\gamma}_K$. Since the exact solution to the problem is not available, the exact errors are not known. For this reason we computed the effectivity indices $\gamma_K = \theta_K / |||e|||_K$, using instead of true errors $|||e|||_K$, the errors understood as a difference between the actual finite element solution and the solution obtained by performing one time step on the mesh enriched to quadratic elements. The global effectivity index for this problem was $\gamma = 7.7$ and a standard deviation of local effectivity indices $\overline{\sigma} = 1.67$.

Example 3: Flat Plate Problem for the Navier-Stokes Equations

The operator splitting algorithm was used to solve a viscous flow past a flat plate. The problem was solved for the following data:

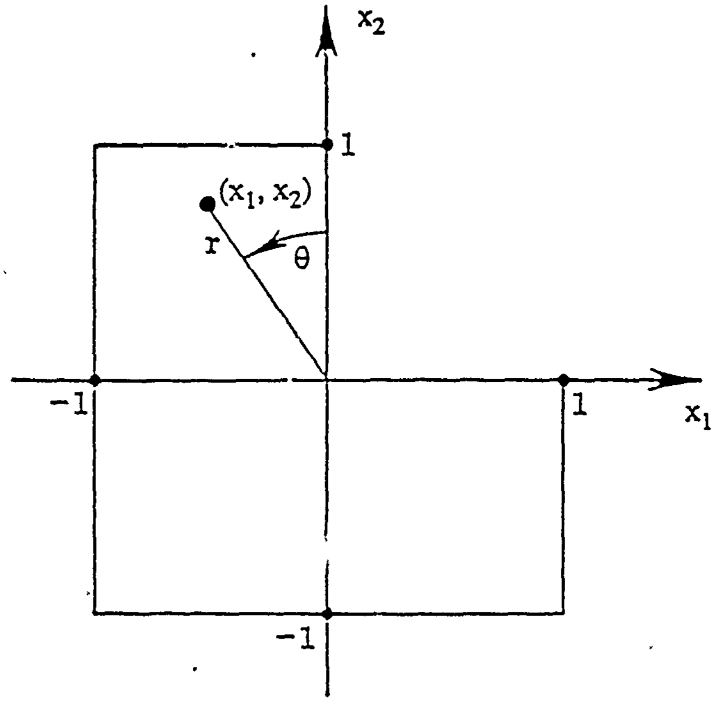
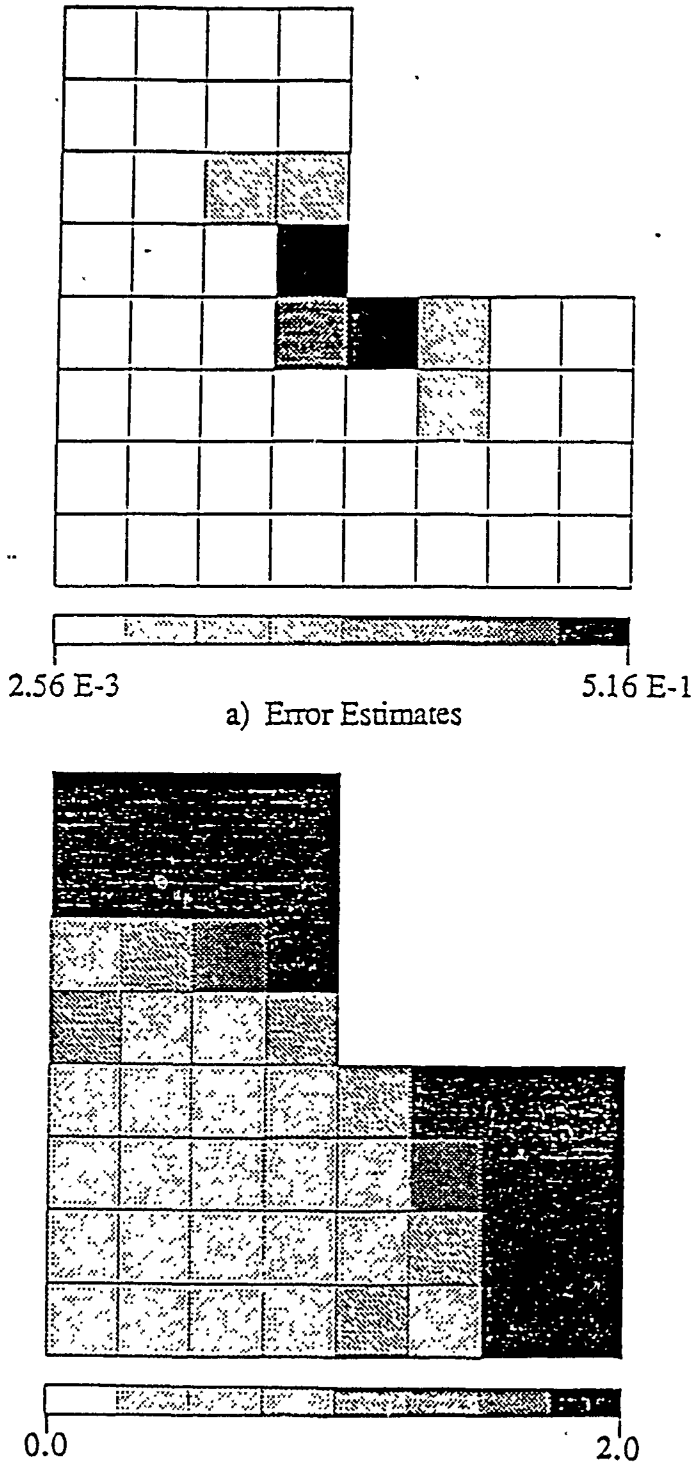


Figure 2: L-shaped domain used for linear diffusion step problem.



a) Error Estimates

b) Normalized Effectivity Indices

Figure 3: (a) Error estimates and (b) normalized effectivity indices for linear diffusion problem. Global effectivity index: $\gamma = 1.276$, standard deviation: $\sigma = 0.537$.

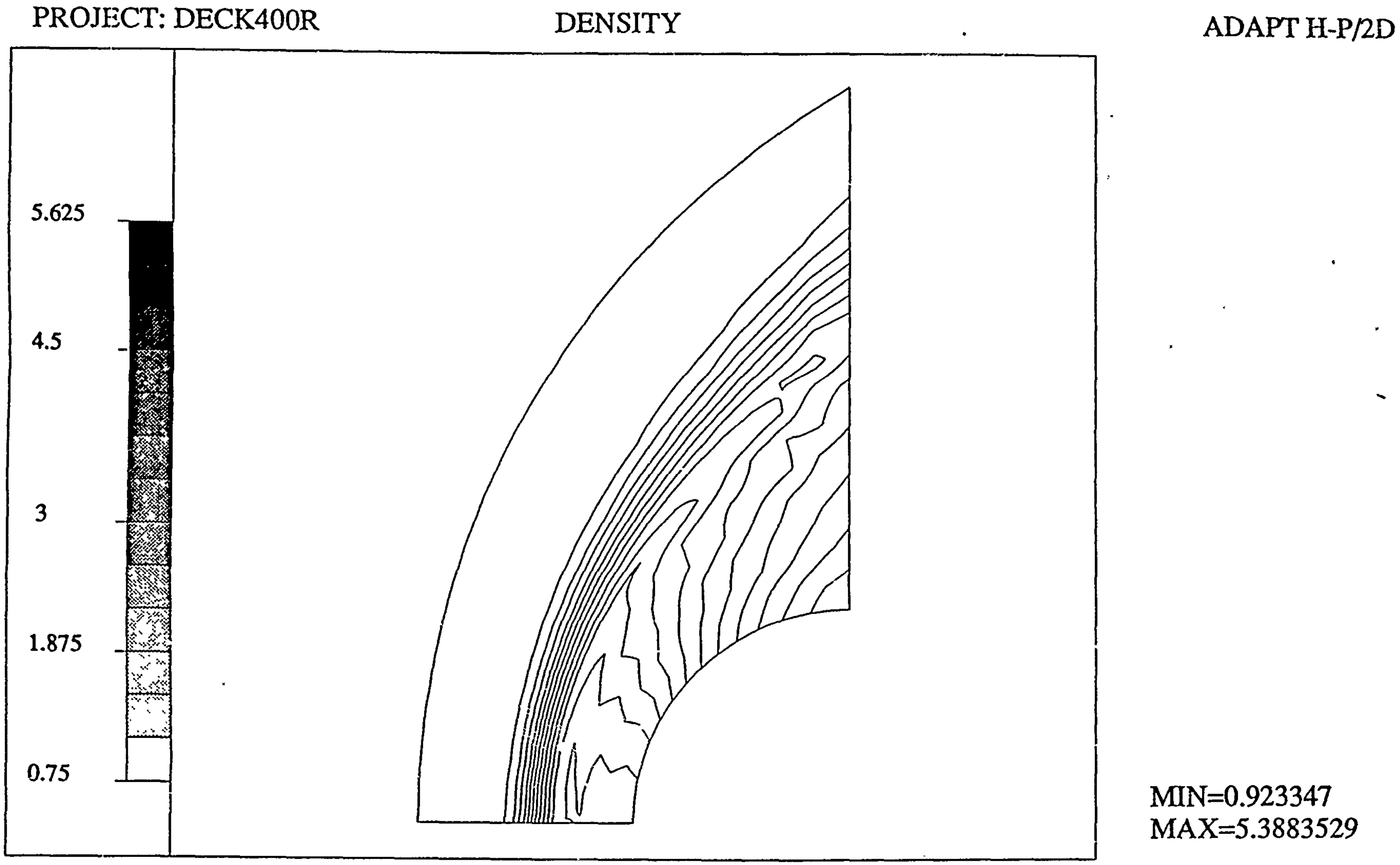


Figure 4: Flow over a blunt body, $Ma = 6$. Density contours.

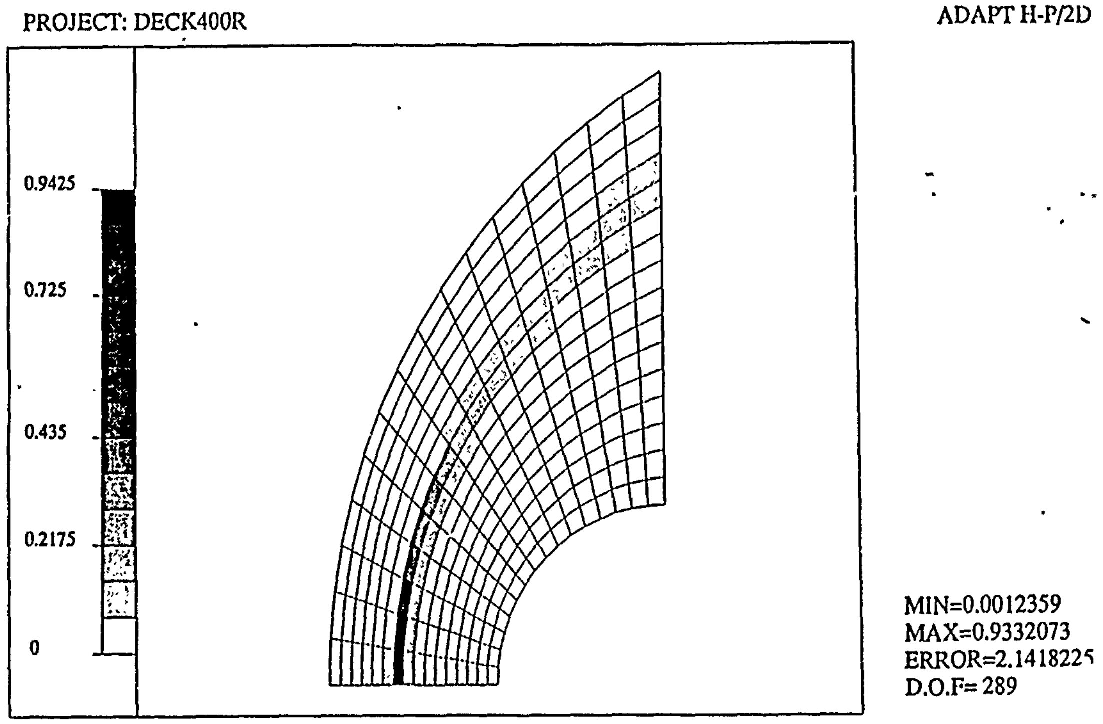


Figure 5: (a) Flow over a blunt body. Distribution of error indicators.

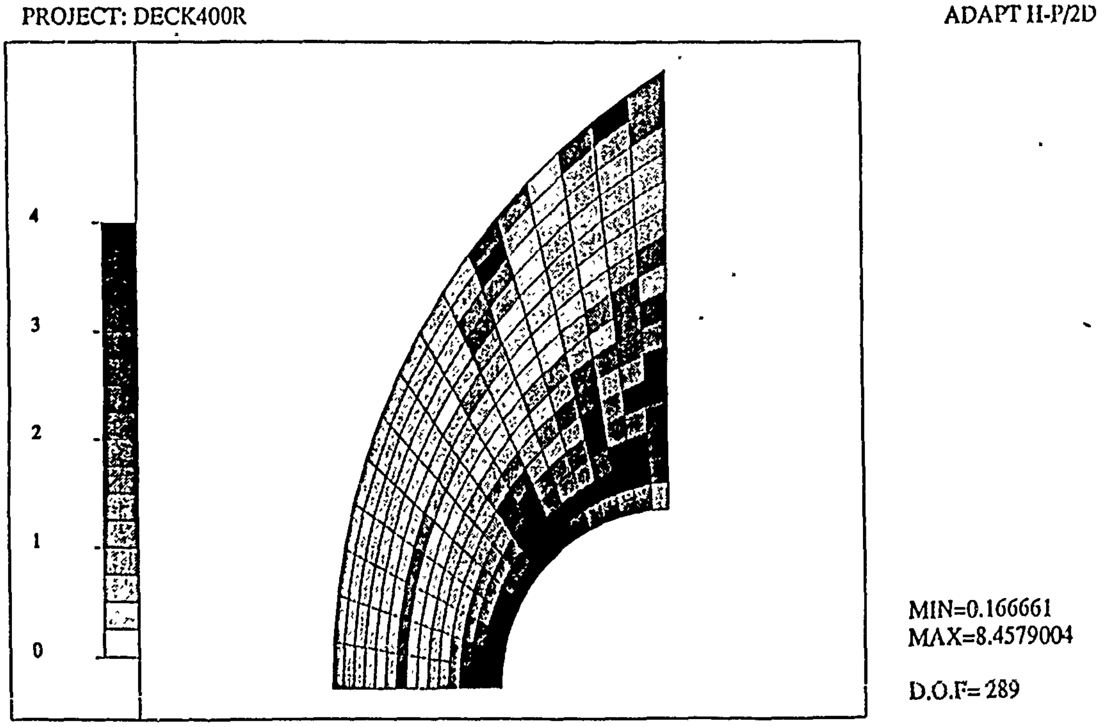


Figure 5: (b) Flow over a blunt body. Local effectivity indices.

- Mach number, $Ma = 3$
- Reynolds number, $Re = 500$
- Free stream temperature $T_\infty = 80°K$
- The temperature of the plate, $T_w = 228°K$

The finite element mesh is shown in Fig. 6. We applied initial h and p refinements to introduce appropriate layers of small higher order (up to $p = 3$) elements along the plate to resolve the boundary layer phenomena. Different shadows of gray in Fig. 6 correspond to different orders of approximation. Elements with only their sides shadowed are anisotropic elements with higher order approximation in the direction perpendicular to the plate only. The solution of the flat plate problem in terms of contours of density is presented in Fig. 7.

Since the viscous splitting algorithm consists of three linear steps, we performed error estimation for all three steps. Similarly, as in Example 2, exact errors involved in effectivity indices analysis were replaced by the errors obtained as differences between the actual solutions of Euler, momentum and energy steps, and the corresponding solutions obtained by enriching the order of approximation by 1 throughout the mesh, and performing one Euler or momentum, or energy time step, respectively. These differences were then measured in energy norms defined by bilinear forms associated with these steps, symmetrized as described in previous sections.

Figures 8, 9, and 10 present distributions of error indicators and local effectivity indices for the three steps of the viscous splitting algorithm. The global effectivity indices γ and standard deviations of local effectivity indices, $\overline{\sigma}$, in this problem were as follows:

Euler step	$\gamma = 18.7$,	$\overline{\sigma} = 6.2$
momentum step	$\gamma = 25.9$,	$\overline{\sigma} = 5.8$
energy step	$\gamma = 3.8$,	$\overline{\sigma} = 7.4$

5 Conclusions

The element residual method has been extended to a general class of symmetrizable variational boundary value problems. The concept has been successfully applied to estimating the error resulting from spatial finite element approximation for steady-state solutions to both the Euler and Navier-Stokes equations obtained using Taylor-Galerkin and operator-splitting methods.

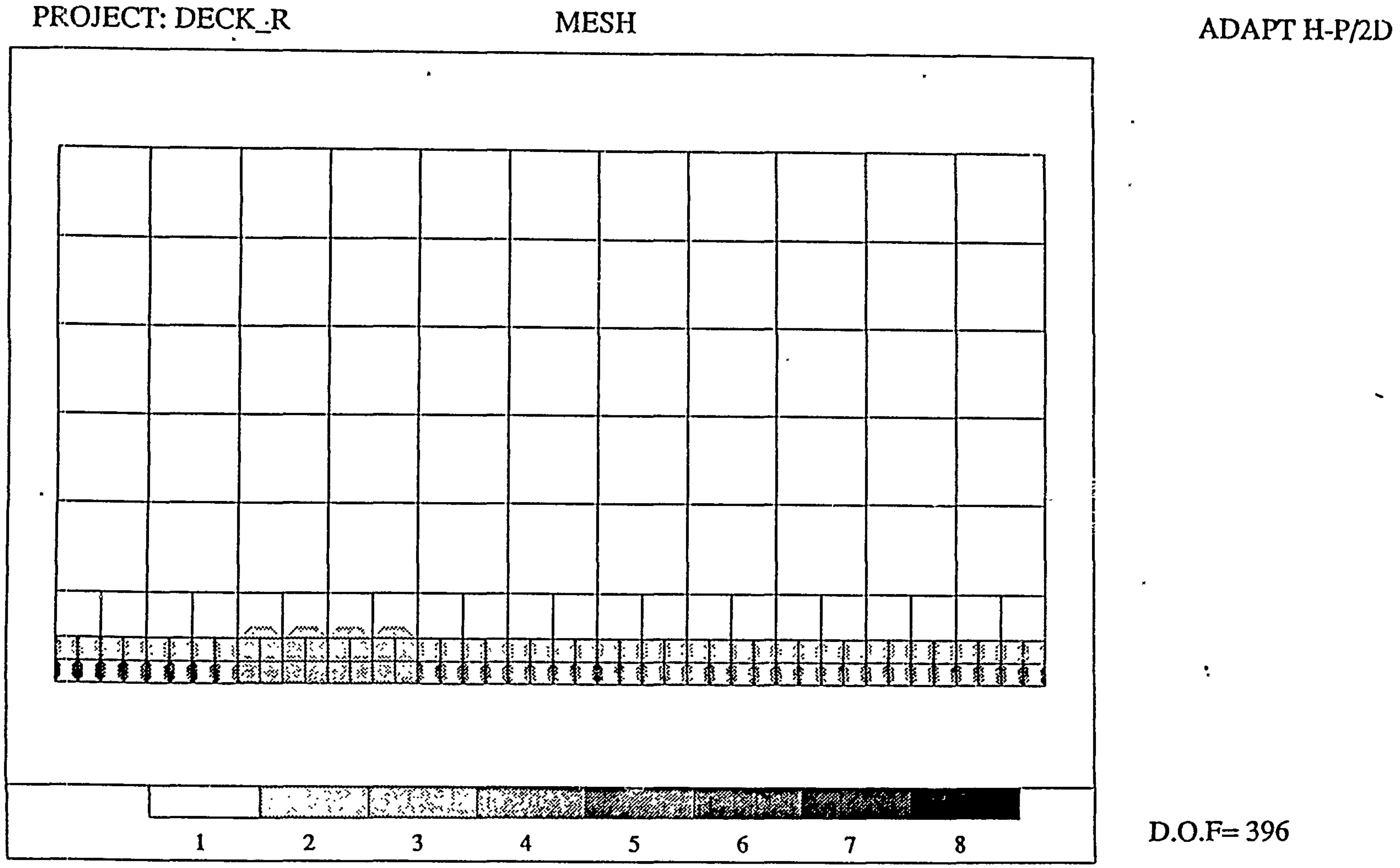


Figure 6: Flat plate problem. An *h-p* finite element mesh.

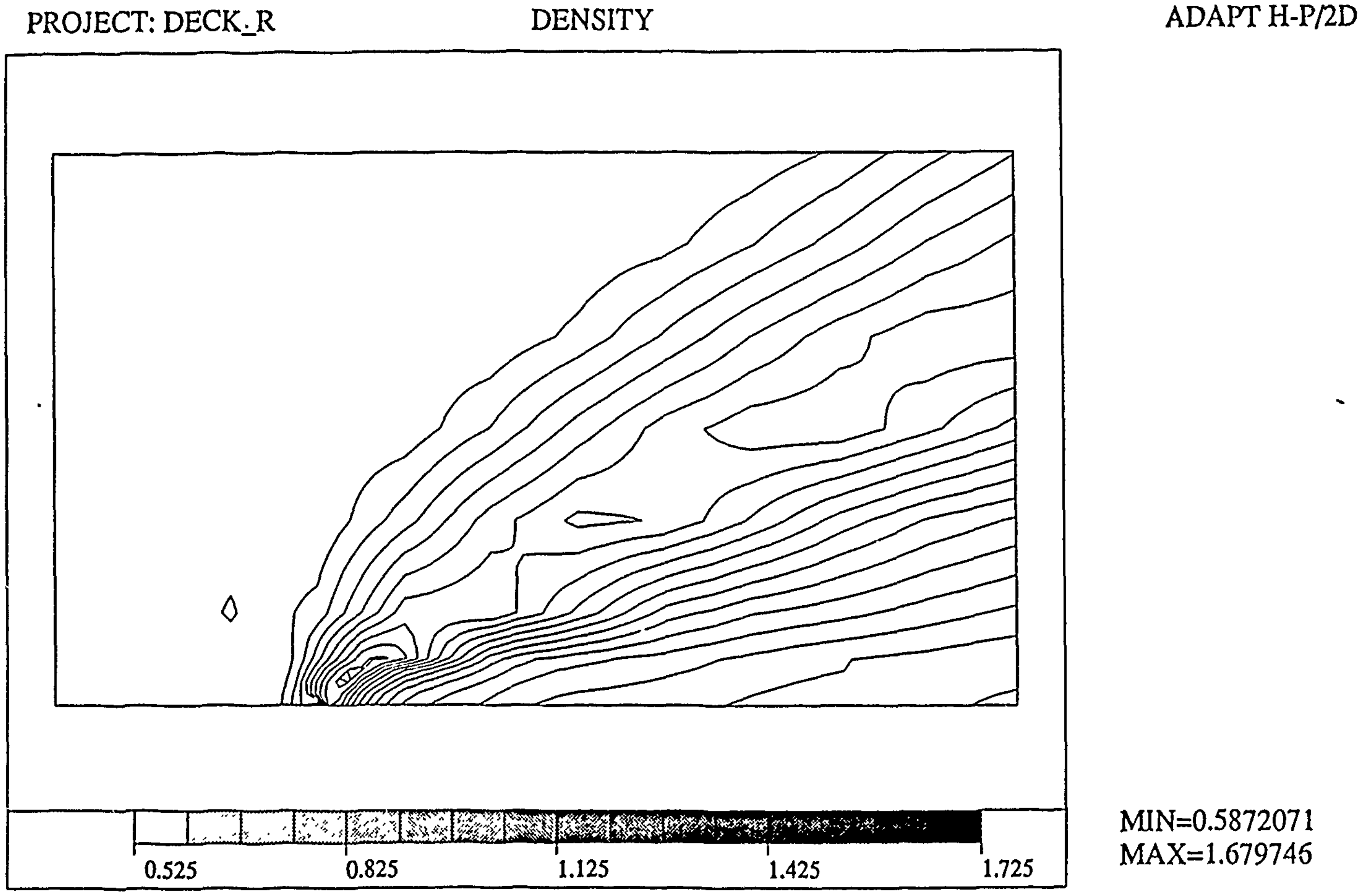


Figure 7: Flat plate problem. Density contours.

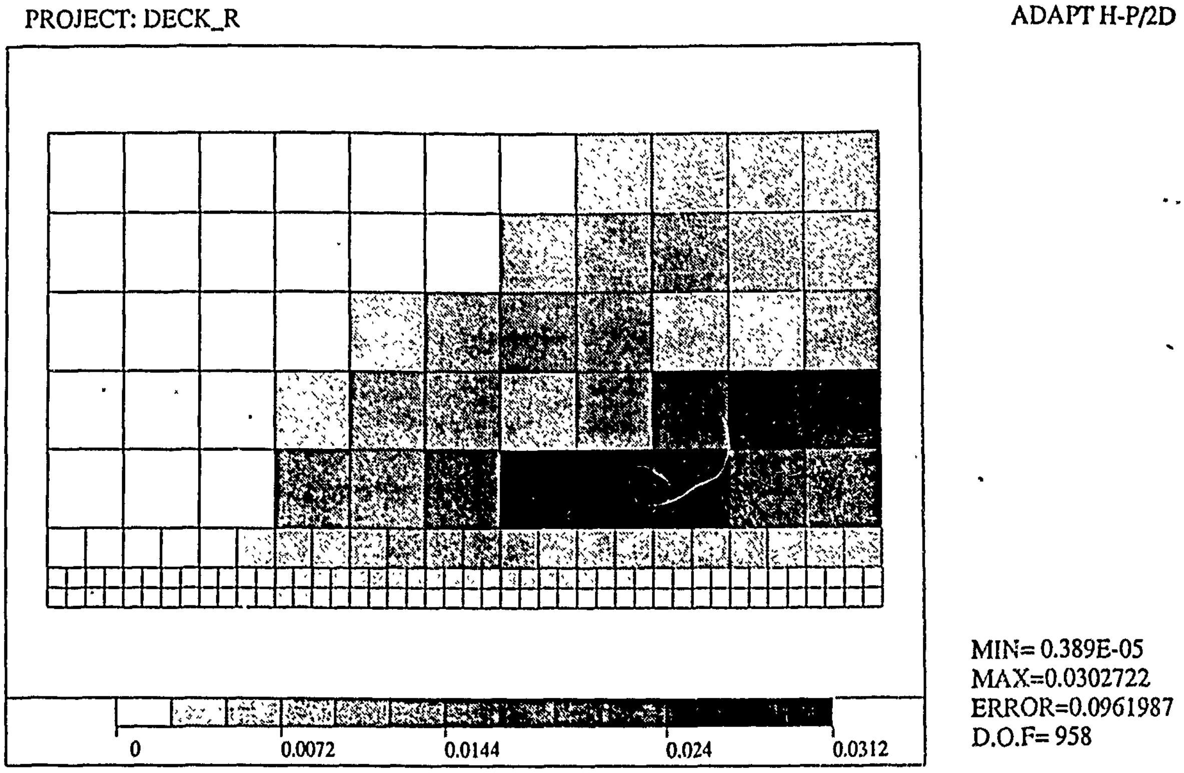


Figure 8: (a) Flat plate problem. Error indicators for the Euler step.

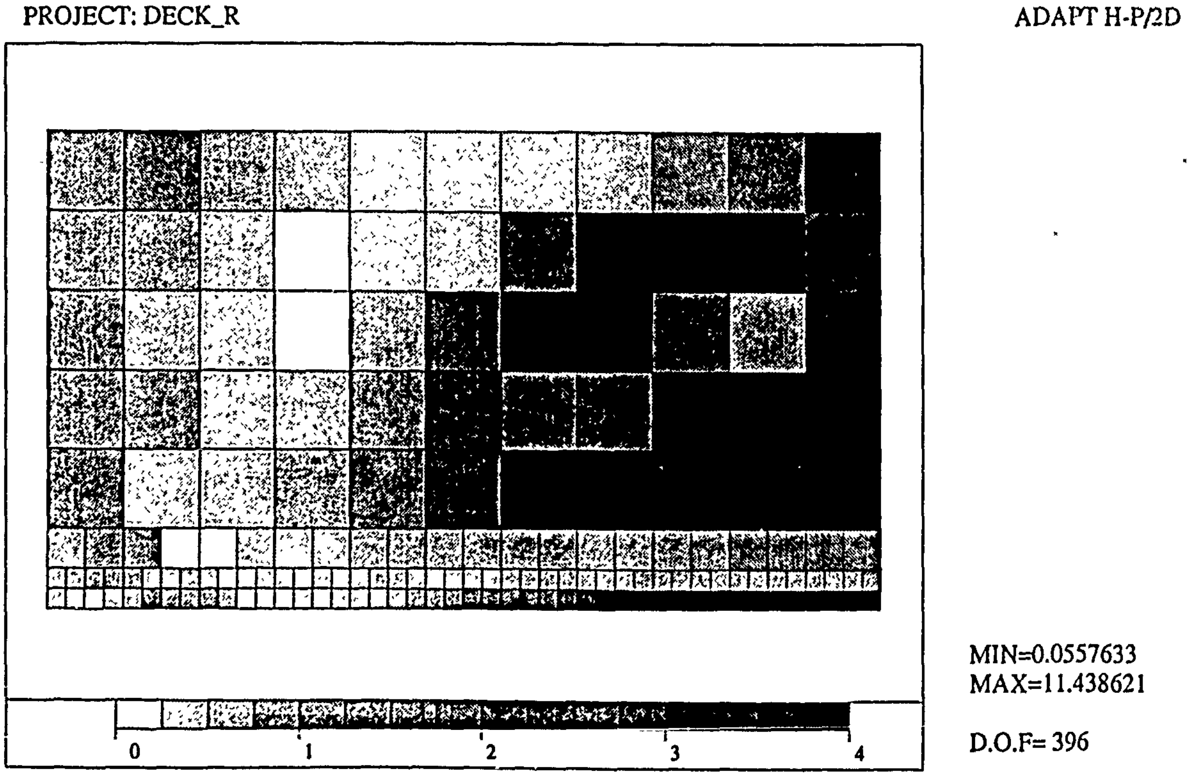


Figure 8: (b) Flat plate problem. Local effectivity indices for the Euler step.

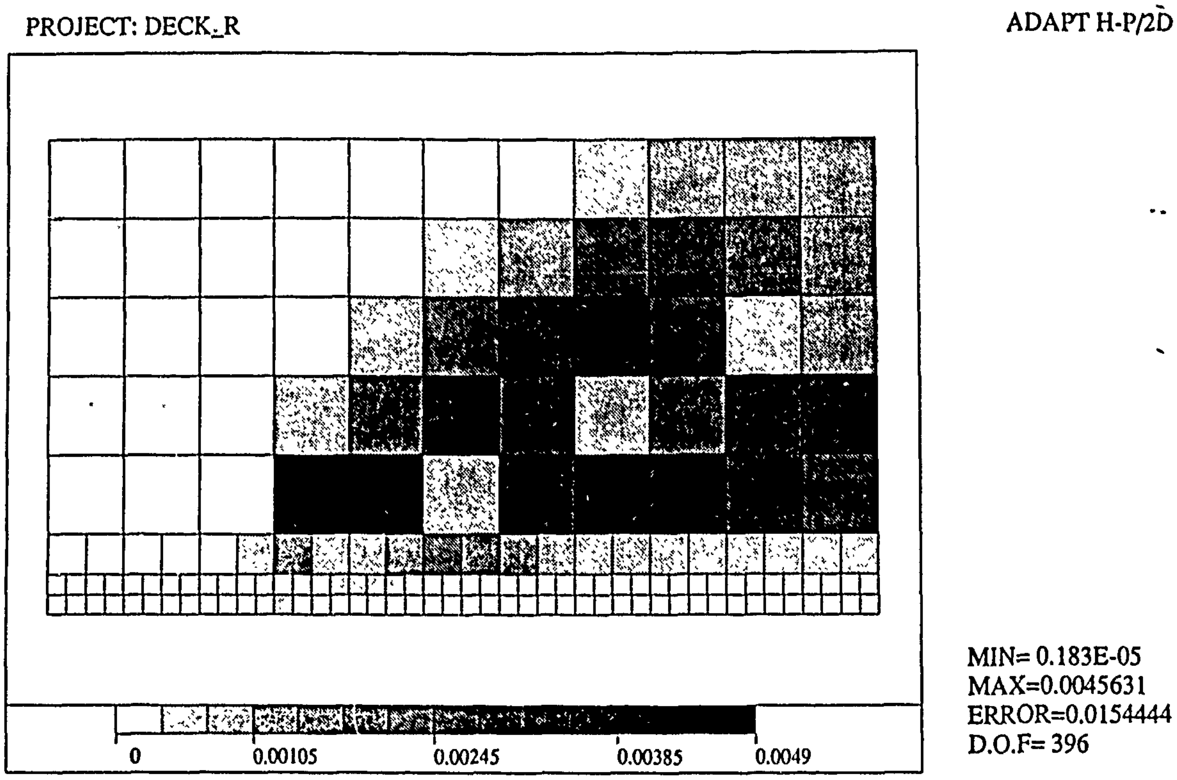


Figure 9: (a) Flat plate problem. Error indicators for the momentum step.

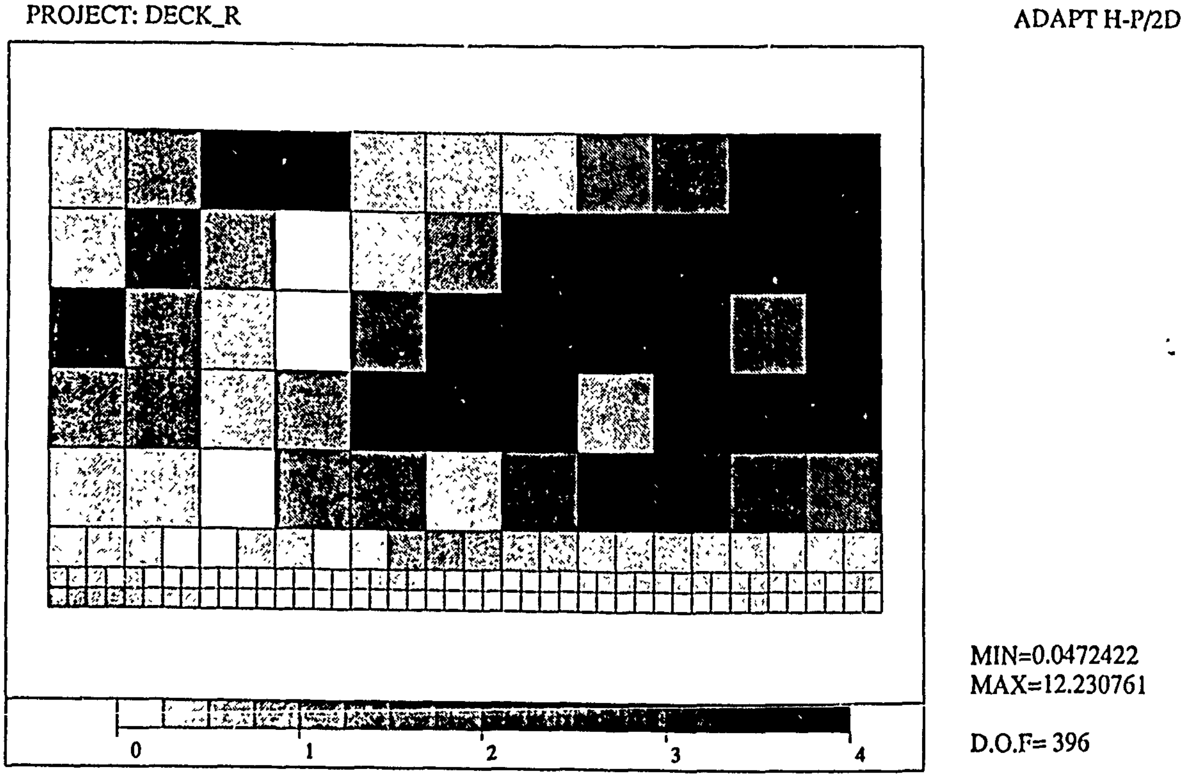


Figure 9: (b) Flat plate problem. Local effectivity indices for the momentum step.

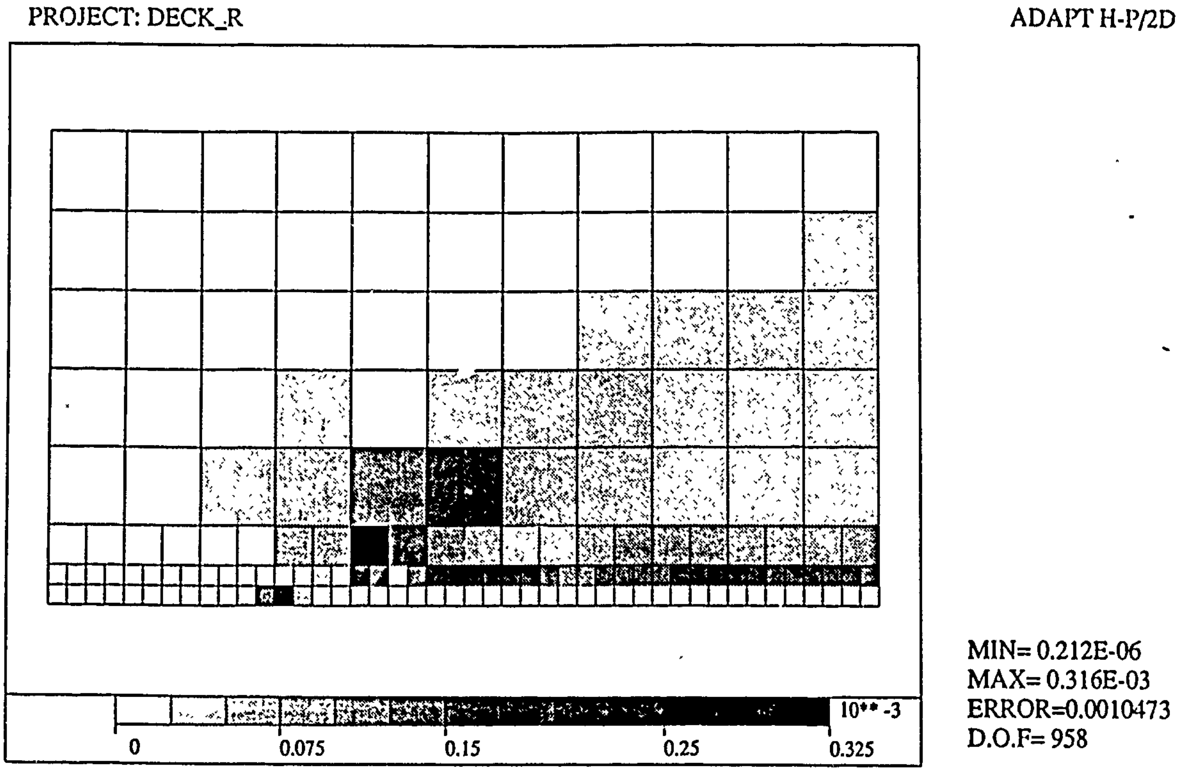


Figure 10: (a) Flat plate problem. Error indicators for the energy step.

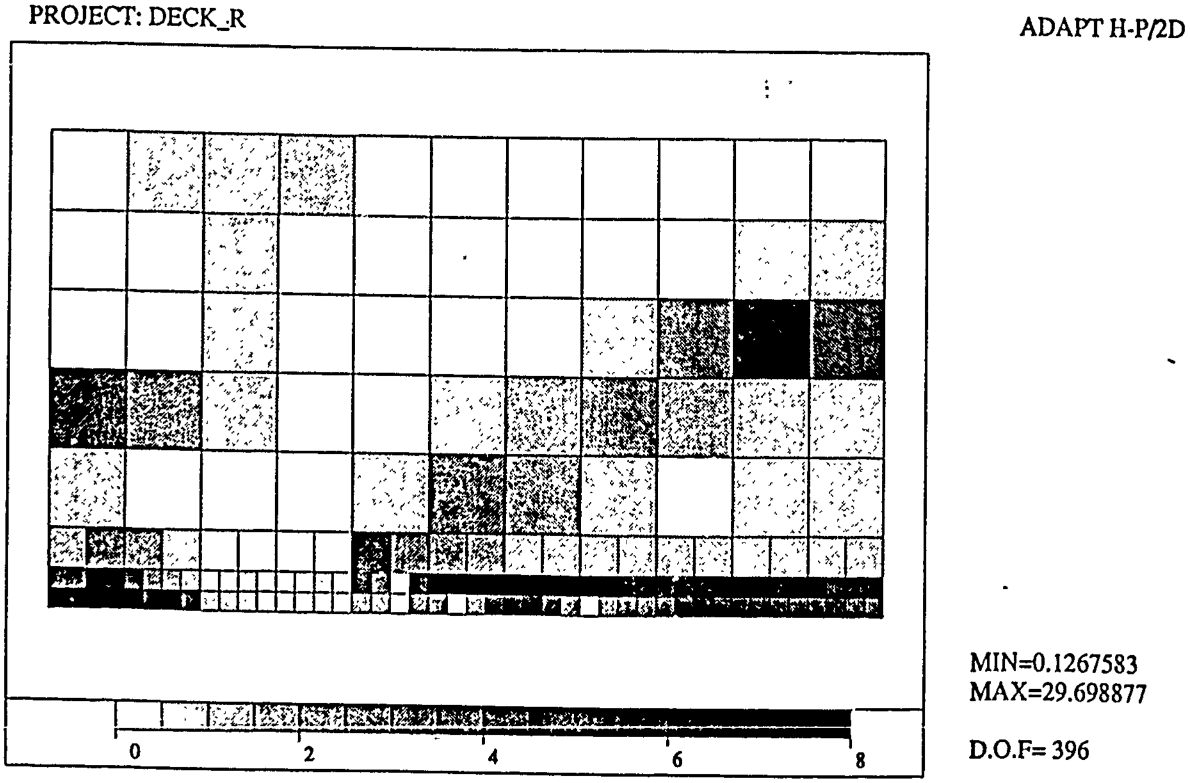


Figure 10: (b) Flat plate problem. Local effectivity indices for the energy step.

Acknowledgement

Our work on error estimation and *h-p* adaptive procedures was initiated under a contract with the Office of Naval Research. Extensions to aerodynamic heating in supersonic flow simulations was done through support of the Aerothermal Loads branch of NASA Langley Research Center. The new operator–splitting techniques, error estimation methods, and anisotropic *p*-refinement methods were developed during a study supported by NASA Ames Research Center.

References

1. Babuška, I., and Rheinboldt, W. C., "*A Posteriori* Error Estimates for the Finite Element Method," *International Journal for Numerical Methods in Engineering*, Vol. 12, pp. 1597–1615, 1978.

2. Bank, R., Welfert, B. D., "*A Posteriori* Error Estimates for the Stokes Equations: A Comparison," submitted to *Computer Methods in Applied Mechanics and Engineering*, in press.

3. Bank, R. E., and Weiser, A., "Some *A Posteriori* Error Estimates for Elliptic Partial Differential Equations," *Mathematics of Computation*, Vol. 44, No. 170, pp. 283–301, 1985.

4. Demkowicz, L., Oden, J. T., and Rachowicz, W., "A New Finite Element Method for Solving Compressible Navier-Stokes Equations Based on an Operator Splitting Method and *h-p* Adaptivity," (in preparation).

5. Demkowicz, L., Oden, J. T., Rachowicz, W., and Hardy, O., "An *h-p* Taylor-Galerkin Finite Element Method for Compressible Euler Equations," (in preparation).

6. Demkowicz, L., Oden, J. T., and Strouboulis, T., "Adaptive Finite Element Methods for Flow Problems With Moving Boundaries. Part I: Variational Principles and *A Posteriori* Estimates," *Computer Methods in Applied Engineering*, Vol. 46, pp. 217–251, 1984.

7. Harten, A., "On the Symmetric Form of Systems of Conservation Laws With Entropy," *Journal of Computational Physics*, Vol. 49, pp. 151–164, 1983.

8. Hughes, T. J. R., Franca, L. P., and Mallet, M., "New Finite Element Formulation for Computational Fluid Dynamics: I. Symetric Forms of the Compressible Euler and

Navier-Stokes Equations and the Second Law of Thermodynamics," *Computer Methods in Applied Mechanics and Engineering*, Vol. 54, pp. 223–234, 1986.

9. Majda, A., "**Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables**," Springer-Verlag, New York 1984.

10. Oden, J. T., and Demkowicz, L., "Advances in Adaptive Improvements: A Survey of Adaptive Finite Element Methods in Computational Mechanics," **State-of-the-Art Surveys in Computational Mechanics**, Edited by A. K. Noor and J. T. Oden, A.S.M.E. Publications, N.Y., 1988.

11. Oden, J. T., Demkowicz, L., Strouboulis, T., and Devloo, Ph., "Adaptive Methods for Problems in Solid and Fluid Mechanics," **Accuracy Estimates and Adaptive Refinements in Finite Element Computations**, Edited by I. Babuška, O. C Zienkiewicz, J., J. Gago, and E. R. de A. Oliveira, John Wiley and Sons, Ltd., Chicnester, pp. 249–280, 1986.

12. Oden, J. T., Demkowicz, L., Rachowicz, W., and Westerman, T. A., "Toward a Universal *h-p* Adaptive Finite Element Strategy, Part 2. *A Posteriori* Error Estimation," *Computer Methods in Applied Mechanics and Engineering*, Vol. pp.

13. Zienkiewicz, O. C., private communication, April 1989.

A Priori Estimates For Mixed Finite Element Methods For The Wave Equation

Lawrence C. Cowsar[1], Todd F. Dupont[2], Mary F. Wheeler[1,3]

Abstract

This paper treats mixed methods for second order hyperbolic equations. The convergence of a mixed method continuous-time scheme for the hyperbolic problem is reduced to a question of convergence of the associated elliptic problem. Stability conditions are also derived for a conditionally stable explicit scheme. Numerical experiments are presented that verify the theoretical rates of convergence and compare two of the discrete schemes discussed.

Keywords: mixed finite element methods, second order hyperbolic equations, stability, convergence

1 Introduction

In a mixed finite element formulation both displacements and stresses are approximated simultaneously. This approach requires the solution of a saddle point problem. These methods provide higher-order approximations of the stresses. This property is important in modeling boundary controllability of the wave equation [10] where accurate forces on the boundary are essential. In computing Darcy velocities in flow in porous media, one is confronted with problems with rough coefficients and anisotropies. Numerical experiments indicate that mixed finite element methods out perform displacement methods [12].

One of the main difficulties of mixed finite element techniques is that convergence and stability require compatibility of the approximating spaces. Discussion of the solvability and stability and the inf-sup condition can be found in other papers [4], [6], [7], [3], even some in this volume, [2].

[1] Department of Mathematics, University of Houston
[2] Department of Computer Science, University of Chicago
[3] Mathematical Sciences Department, Rice University

In this paper we establish *a priori* convergence results for continuous-time mixed finite element methods for second order hyperbolic problems. We address what we feel are essential theoretical results. More precisely, we show that it is possible to reduce the question of convergence of the continuous-time scheme to the question of convergence of the mixed method for the associated elliptic problem. In addition we discuss stability of a collection of time-stepping schemes. A time-step condition for stability is exhibited for an explicit procedure. Approximation of initial conditions and boundary conditions is discussed.

A priori error estimates for Galerkin approximations for second order hyperbolic equations have been previously derived by Dupont [11] and improved by Baker [5]. Both continuous and discrete time schemes were analyzed. The estimates of Dupont were based on the usual energy inequality for the wave equation, and those of Baker can be viewed as being based on a nonstandard type of energy relation.

In [13] Geveci considers mixed finite element methods for the wave equation. He replaced the wave equation by a first order system, both in space and time. The continuous-time procedure discussed here is equivalent to the one he treats. The error estimates presented here are analogues of the improved estimates of Baker, while the error bounds of Geveci parallel the continuous-time results of Dupont. The discrete time methods presented here are different from those in Geveci's paper.

The paper consists of four additional sections. In Section 2, mixed finite element methods are formulated for both continuous and discrete time. A convergence result for the continuous-time scheme is established in Section 3. Stability for the discrete time schemes is discussed in Section 4. In particular we derive a stability condition for the explicit formulation. Numerical experiments are presented in Section 5.

2 Mixed Finite Element Formulation for Second Order Hyperbolic Equations

We consider the second order hyperbolic equation

$$(1)\quad \frac{\partial^2 u}{\partial t^2} - \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(\mathrm{x})\frac{\partial u}{\partial x_j}(\mathrm{x},t)\right) = f(\mathrm{x},t),\ \mathrm{x}\in\Omega,\ \ t\in(0,T],$$

with initial conditions,

$$u(\mathrm{x},0)=u_0(\mathrm{x}), \qquad \mathrm{x}\in\Omega, \tag{2}$$

$$\frac{\partial u}{\partial t}(\mathrm{x},0)=u_1(\mathrm{x}), \qquad \mathrm{x}\in\Omega, \tag{3}$$

where Ω is a bounded domain in R^n with boundary $\partial\Omega$. For convenience we assume homogeneous Dirichlet boundary conditions:

$$u(\mathrm{x},t)=0, \quad \mathrm{x}\in\partial\Omega,\ t>0. \tag{4}$$

We assume that the functions a_{ij} and f are uniformly bounded and measurable. In addition we assume that the spatial operator is uniformly elliptic in the sense that there exists positive constants α and β such that

$$\alpha\sum_{i=1}^{n}\xi_i^2\le\sum_{i,j=1}^{n}a_{ij}(\mathrm{x})\xi_i\xi_j\le\beta\sum_{i=1}^{n}\xi_i^2, \quad \mathrm{x}\in\bar{\Omega},\ \xi\in R^n. \tag{5}$$

We adopt the following notation: Let $H(\Omega;\mathrm{div})$ be the subspace of $(L^2(\Omega))^n$ defined by

$$H(\Omega;\mathrm{div})=\{\mathrm{z}|\mathrm{div}\ \mathrm{z}\in L^2(\Omega)\}. \tag{6}$$

Let $(\cdot,\cdot)$ denote the L^2 inner product,

$$(\varphi,\psi)\equiv\int_\Omega\phi(\mathrm{x})\psi(\mathrm{x})d\mathrm{x}, \ \ \psi,\phi\in L^2(\Omega), \tag{7}$$

$$\|\phi\|=(\phi,\phi)^{\frac{1}{2}}. \tag{8}$$

For s = 0, 1, ... denote by $H^s(\Omega)$ the usual Sobolev space of real-valued functions with s derivatives in $L^2(\Omega)$. Also $H^{-1}(\Omega)$ is defined as the completion of $C^\infty(\Omega)$ with respect to the norm

$$\|\phi\|_{-1}=\sup_{\substack{\psi\in C^\infty(\Omega)\\ \psi\neq 0}}\frac{|(\phi,\psi)|}{\|\psi\|_1}, \tag{9}$$

For definitions and the relevant properties of these spaces, we refer to [1], [14], [17].

For H, a normed space with norm $\|\cdot\|_H$ and $\phi:[0,T]\to H$ sufficiently regular, the following norms are defined:

$$\text{(10)} \qquad \| \phi \|^p_{L^p((0,T);H)} = \int_0^T \| \phi(\cdot,t) \|^p_H \, dt, \ \ 1 \le p < \infty,$$

$$\text{(11)} \qquad \| \phi \|_{L^\infty((0,T);H)} = \sup_{0 \le t \le T} \| \phi(\cdot,t) \|_H \, .$$

Before defining a mixed finite element procedure we rewrite (1) in the following weak formulation [9]

$$\text{(12)} \qquad (\frac{\partial^2 u}{\partial t^2}, \chi) + (\text{div } z, \chi) \ = (f, \chi), \qquad \chi \in L^2(\Omega),$$

$$\text{(13)} \qquad (A^{-1} z, v) - (u, \text{div} v) \ = 0, \quad v \in H(\Omega, \text{div}),$$

where

$$\text{(14)} \qquad z = -A \nabla u.$$

Here we derive a nonstandard "energy inequality" for the solution of (1). This energy inequality is motivated by Baker's work [5] and is used in the analysis of the mixed method. It also sheds light on the proper choice of the initial conditions. Define

$$\text{(15)} \qquad \phi(\mathrm{x}, t) \ = \ \int_0^t u(\mathrm{x}, s) ds,$$

$$\text{(16)} \qquad \psi(\mathrm{x}, t) \ = \ \int_0^t z(\mathrm{x}, s) ds.$$

Thus,

$$\text{(17)} \qquad \frac{\partial^2 \phi}{\partial t^2} \ = \ \frac{\partial u}{\partial t}$$

$$\text{(18)} \qquad \ = \ \int_0^t \frac{\partial^2 u}{\partial t^2} + u_1.$$

Hence

$$\text{(19)} \qquad \begin{aligned} \frac{\partial^2 \phi}{\partial t^2} - \nabla \cdot A \nabla \phi \ &= \ \int_0^t (\frac{\partial^2 u}{\partial t^2}(\mathrm{x}, s) - \nabla \cdot A \nabla u) ds + u_1 \\ &= \ \int_0^t f(\mathrm{x}, s) + u_1 \\ &\equiv \ \theta + u_1. \end{aligned}$$

Note that (12) and (13) become

$$\text{(20)} \qquad (\frac{\partial^2 \phi}{\partial t^2}, \chi) + (\text{div} \psi, \chi) \ = \ (\theta, \chi) + (u_1, \chi), \qquad \chi \in L^2(\Omega),$$

$$\text{(21)} \qquad (A^{-1} \psi, v) - (\phi, \text{div} v) \ = \ 0, \quad v \in H(\Omega; \text{div}).$$

Taking the L^2 inner product of (19) with $\frac{\partial\phi}{\partial t}$ gives

$$
\begin{aligned}
(22)\qquad \frac{1}{2}\frac{d}{dt}\parallel \frac{\partial\phi}{\partial t}\parallel^2 + \frac{1}{2}\frac{d}{dt}\parallel A^{\frac{1}{2}}\nabla\phi\parallel^2 &= (\theta + u_1, \frac{\partial\phi}{\partial t}) \\
&= (\theta, \frac{\partial\phi}{\partial t}) + \frac{d}{dt}(u_1, \phi).
\end{aligned}
$$

Integrating (22) from 0 to T, we obtain

$$
\begin{aligned}
(23)\qquad &\parallel \frac{\partial\phi}{\partial t}\parallel^2 (T) + \parallel A^{\frac{1}{2}}\nabla\phi\parallel^2 (T) = \parallel \frac{\partial\phi}{\partial t}\parallel^2 (0) + \parallel A^{\frac{1}{2}}\nabla\phi\parallel^2 (0) \\
&+2\left[-\int_0^T (\frac{\partial\theta}{\partial t}(\cdot,t), \phi(\cdot,t)) + (\theta(\cdot,T), \phi(\cdot,T)) + (u_1, \phi(\cdot,T))\right] \\
&= \parallel u_0\parallel^2 + 2[-\int_0^T (f(\cdot,t), \phi(\cdot,t)) + (\theta(\cdot,T), \phi(\cdot,T)) + (u_1, \phi(\cdot,T))].
\end{aligned}
$$

By the Schwartz inequality and duality, the right hand side of (23) is bounded by

$$
\begin{aligned}
(24)\qquad & C(\alpha)(\parallel u_0\parallel^2 + \parallel u_1\parallel^2_{-1} + \parallel f\parallel^2_{L^2((0,T);H^{-1})}) \\
+ \ & \frac{1}{4}\parallel A^{\frac{1}{2}}\nabla\phi\parallel^2 (T) + \parallel A^{\frac{1}{2}}\nabla\phi\parallel_{L^2((0,T),L^2)} .
\end{aligned}
$$

Combining (15), (23), and (24) and applying Gronwall's Lemma, we deduce the following:

$$
\begin{aligned}
(25)\qquad \parallel u\parallel^2 (T) &\le \parallel \frac{\partial\phi}{\partial t}\parallel^2 (T) + \parallel A^{\frac{1}{2}}\nabla\phi\parallel^2 (T) \\
&\le C(\alpha)(\parallel u_0\parallel^2 + \parallel u_1\parallel^2_{-1} + \parallel f\parallel^2_{L^2((0,T);H^{-1})}).
\end{aligned}
$$

For h a small positive parameter we take W_h and V_h to be finite dimensional subspaces of $L^2(\Omega)$ and $H(\Omega, \text{div})$, respectively. These spaces will need to satisfy some compatibility constraints. However, for our purposes we merely suppose that the mixed finite element approximation of elliptic problems is well posed for these spaces.

The continuous-time mixed finite element approximation to (1)-(4) is defined as a map from $[0,T]$ into $W_h \times V_h$ given by the pair $(U(\cdot,t), Z(\cdot,t))$ satisfying

$$
\begin{aligned}
(26)\qquad (\frac{\partial^2 U}{\partial t^2}, \chi) + (\text{div}Z, \chi) &= (f, \chi), \quad \chi \in W_h, \quad t > 0, \\
(27)\qquad (A^{-1}Z, v) - (U, \text{div}v) &= 0, \quad v \in V_h, \quad t > 0.
\end{aligned}
$$

We define $U(\cdot,0)$ and $\frac{\partial U}{\partial t}(\cdot,0)$ by

$$(U(\cdot,0)-u_0,w)=0, \qquad w\in W_h, \tag{28}$$

$$\left(\frac{\partial U}{\partial t}(\cdot,0)-u_1,w\right)=0, \qquad w\in W_h. \tag{29}$$

A family of discrete time mixed finite element procedures can be defined as follows.

Let $\Delta t>0$ and $t_n=n\Delta t$. Define

$$\phi^n \;=\; \phi(\cdot,t^n), \tag{30}$$

$$\partial_t\phi^{n+\frac{1}{2}} \;=\; (\phi^{n+1}-\phi^n)/\Delta t, \tag{31}$$

$$\partial_t^2\phi^n \;=\; (\phi^{n+1}-2\phi^n+\phi^{n-1})/(\Delta t)^2, \tag{32}$$

$$\phi^{n,s} \;=\; s\phi^{n+1}+(1-2s)\phi^n+s\phi^{n-1}. \tag{33}$$

For $0\le s\le 1$, the s-discrete mixed finite element method approximation $U^n\in W_h$, $Z^n\in V_h$ is defined by

$$\left(\frac{U^1-U^{-1}}{2\Delta t},\chi\right) \;=\; (u_1,\chi),\ \chi\in W_h, \tag{34}$$

$$(U^0,\chi) \;=\; (u_0,\chi),\ \ \chi\in W_h, \tag{35}$$

$$(\partial_t^2U^n,\chi)+(\mathrm{div}Z^{n,s},\chi) \;=\; (f^{n,s},\chi),\ \chi\in W_h, n\ge 0, \tag{36}$$

$$(A^{-1}Z^{n,s},v)-(U^{n,s},\mathrm{div}v) \;=\; 0, v\in V_h,\ n\ge -1. \tag{37}$$

For $s=0$ existence of the solution to this system is clear, but for nonzero s the existence of the solution follows from the fact that one can solve the mixed method problem for the elliptic operator

$$-\,\mathrm{div}A\nabla u+\frac{1}{2\ s\ \Delta t^2}u. \tag{38}$$

3 Convergence of the Continuous-Time Mixed Finite Element Method

In this section we derive *a priori* error estimates for the numerical approximation (26)-(27) to (1). In particular we show that convergence of the transient second order hyperbolic problem follows from convergence theory for mixed finite element methods for elliptic problems. Define

$$\Phi(\mathrm{x},t) = \int_0^t U(\mathrm{x},s)ds, \tag{39}$$

$$\Psi(\mathrm{x},t) = \int_0^t Z(\mathrm{x},s)ds, \tag{40}$$

where $U(\cdot,t)$ and $Z(\cdot,t)$ by (19) and (24).
We observe that

$$\begin{aligned}
\frac{\partial \Phi}{\partial t} &= U(\mathrm{x},t), & (41)\\
\frac{\partial^2 \Phi}{\partial t^2} &= \frac{\partial}{\partial t}U(\mathrm{x},t) = \int_0^t \frac{\partial^2 U}{\partial t^2} ds + U(\cdot,0). & (42)
\end{aligned}$$

Integrating (26) and (27) from 0 to t and using (39) and (40) we obtain

$$\begin{aligned}
\left(\frac{\partial^2 \Phi}{\partial t^2},\chi\right) + (\mathrm{div}\Psi,\chi) &= (\theta,\chi) + (u_1,\chi), \quad \chi \in W_h, & (43)\\
(A^{-1}\Psi, v) - (\Phi, \mathrm{div} v) &= 0, \quad v \in V_h, & (44)
\end{aligned}$$

where θ is given by (19).

We compare the pair $(\Phi(\cdot,t),\Psi(\cdot,t))$ with $(\tilde{\Phi}(\cdot,t)),\tilde{\Psi}(\cdot,t)) \in W_h \times V_h$ defined by

$$(\mathrm{div}(\tilde{\Phi} - \phi),\chi) = 0, \quad \chi \in W_h, \tag{45}$$

$$(A^{-1}(\tilde{\Phi} - \phi), v) - (\tilde{\Psi} - \psi, \mathrm{div} v) = 0, \quad v \in V_h, \tag{46}$$

for $t \in [0,T]$, where (ϕ,ψ) are given by (15) and (16). Using (20), (21), (45) and (46) we observe that

$$\left(\frac{\partial^2 \tilde{\Phi}}{\partial t^2},\chi\right) + (\mathrm{div}\tilde{\Psi},\chi) = (\theta,\chi) + \left(\frac{\partial^2 \eta}{\partial t^2},\chi\right) + (u_1,\chi), \quad \chi \in W_h, \tag{47}$$

$$(A^{-1}\tilde{\Psi}, v) - (\tilde{\Phi}, \mathrm{div} v) = 0, \quad v \in V_h, \tag{48}$$

where $\eta = \phi - \tilde{\Phi}$.

We also note that $\tilde{U} = \frac{\partial}{\partial t}\tilde{\Phi}$ and $\tilde{Z} = \frac{\partial}{\partial t}\tilde{\Psi}$ satisfy

$$(\mathrm{div}(\tilde{Z} - z),\chi) = 0, \quad \chi \in W_h, \tag{49}$$

$$(A^{-1}(\tilde{Z} - z), v) - ((\tilde{U} - u), \mathrm{div} v) = 0, \quad v \in V_h. \tag{50}$$

We see that $(\tilde{U}, \tilde{Z})$ is exactly the elliptic finite element approximation of (u, z).

Set

$$\Gamma = \Phi - \tilde{\Phi} \quad \in W_h, \tag{51}$$

$$\mu = \Psi - \tilde{\Psi} \quad \in V_h. \tag{52}$$

Subtracting (47) and (48) from (43) and (44) respectively we have

$$\left(\frac{\partial^2\Gamma}{\partial t^2}, \chi\right) + (\mathrm{div}\mu, \chi) = -\left(\frac{\partial^2\eta}{\partial t^2}, \chi\right), \quad \chi \in W_h, \tag{53}$$

$$(A^{-1}\mu, v) - (\Gamma, \mathrm{div} v) = 0 \quad , \quad v \in V_h. \tag{54}$$

Differentiating (54) with respect to t, we see that

$$\left(A^{-1}\frac{\partial\mu}{\partial t}, v\right) - \left(\frac{\partial\Gamma}{\partial t}, \mathrm{div} v\right) = 0, \quad v \in V_h. \tag{55}$$

Combining (53) with $\chi = \frac{\partial\Gamma}{\partial t}$ and (55) with $v = \mu$ we have

$$\frac{1}{2}\frac{d}{dt} \| \frac{\partial\Gamma}{\partial t} \|^2 + \frac{1}{2}\frac{d}{dt}(A^{-1}\mu, \mu) = -\left(\frac{\partial^2\eta}{\partial t^2}, \frac{\partial\Gamma}{\partial t}\right). \tag{56}$$

Integrating (56) from 0 to T and noting that $\mu(0) = 0$, we deduce that

$$\begin{aligned} \| \frac{\partial\Gamma}{\partial t} \|^2 (T) + \| A^{-\frac{1}{2}}\mu \|^2 (T) & \\ \leq \| \frac{\partial\Gamma}{\partial t} \|^2 (0) + 2\int_0^T \| \frac{\partial^2\eta}{\partial t^2} \| \| \frac{\partial\Gamma}{\partial t} \| & \\ \leq \| \frac{\partial\Gamma}{\partial t} \|^2 (0) + 2 \| \frac{\partial^2\eta}{\partial t^2} \|^2_{L^1((0,T);L^2)} + \frac{1}{2} \| \frac{\partial\Gamma}{\partial t} \|^2_{L^\infty((0,T);L^2)} & \end{aligned} \tag{57}$$

Now

$$\| \frac{\partial\Gamma}{\partial t} \| (0) \leq \| \frac{\partial\eta}{\partial t} \| (0). \tag{58}$$

Considering (57) and (58) we see that

$$\begin{aligned} \| \frac{\partial\Gamma}{\partial t} \|^2_{L^\infty((0,T);L^2)} + \| A^{-\frac{1}{2}}\mu \|^2_{L^\infty((0,T);L^2)} & \\ \leq 4 \left(\| \frac{\partial^2\eta}{\partial t^2} \|^2_{L^1((0,T);L^2)} + \| \frac{\partial\eta}{\partial t} \|^2 (0) \right). & \end{aligned} \tag{59}$$

Using (59) and the triangle inequality, one can easily deduce that

$$(60)\qquad \| U - u \|_{L^\infty((0,T);L^2)} \leq 3 \left[\| u - \tilde{U} \|_{L^\infty((0,T);L^2)} + \| \frac{\partial(u - \tilde{U})}{\partial t} \|_{L^1((0,T);L^2)} \right].$$

Theorem 1 *Let (U, Z) denote the continuous-time mixed finite element approximation given by (26)-(29) which approximates the second order hyperbolic problem (1)-(4). Then the errors are bounded by (60) where $(\tilde{U}, \tilde{Z})$ are the elliptic mixed finite element approximations given by (49)-(50).*

This result shows that a convergence result for the elliptic procedure gives a corresponding convergence result for the second order hyperbolic process. Note that the error bound (60) requires that the elliptic mixed method approximate $\frac{\partial u}{\partial t}$ well, and in this sense the bound requires more smoothness than would be required by approximation theory.

4 Stability of the Discrete-Time Mixed Finite Element Approximations

In this section we derive a stability estimate for the explicit scheme $s = 0$ in (34)-(37). We suppose the following "inverse assumption": There exists a constant C_0, independent of h, such that

$$(61)\qquad \| \text{div}\phi \| \leq C_0 h^{-1} \| \phi \|, \quad \phi \in V_h.$$

Specifically, we have

$$(62)\qquad (\partial_t^2 U^n, w) + (\text{div} Z^n, w) = 0, \quad w \in W_h,$$

$$(63)\qquad (A^{-1} Z^n, v) - (U^n, \text{div} v) = 0, \quad v \in V_h.$$

Subtracting (63) from itself with n replaced by $n + 1$ and $n - 1$ respectively, we have

$$(64)\qquad (A^{-1}(Z^{n+1} - Z^{n-1}), v) - (U^{n+1} - U^{n-1}, \text{div} v) = 0, \ v \in V_h.$$

Setting $w = \frac{(U^{n+1} - U^{n-1})}{2\Delta t}$ in (62) and $v = \frac{Z^n}{2\Delta t}$ in (64) and adding the two equations, we obtain

$$(\partial_t^2 U^n, \frac{U^{n+1} - U^{n-1}}{2\Delta t}) + (A^{-1}(\frac{Z^{n+1} - Z^{n-1}}{2\Delta t}), Z^n) = 0.$$

Now from (31) we have

$$\partial_t^2 U^n = \frac{((U^{n+1}-U^n)-(U^n-U^{n-1}))}{(\Delta t)^2} \tag{65}$$

$$= \frac{(\partial_t U^{n+\frac{1}{2}}-\partial_t U^{n-\frac{1}{2}})}{\Delta t}. \tag{66}$$

Now,

$$\delta_t U^n = \frac{(U^{n+1}-U^{n-1})}{2\Delta t} \tag{67}$$

$$= \frac{1}{2}(\partial_t U^{n+\frac{1}{2}} + \partial_t U^{n-\frac{1}{2}}). \tag{68}$$

Thus,

$$\begin{aligned} &(\partial_t^2 U^n, \delta_t U^n) + (A^{-1}\delta_t Z^n, Z^n) \\ &= \frac{1}{2\Delta t}(\partial_t U^{n+\frac{1}{2}} - \partial_t U^{n-\frac{1}{2}}, \partial_t U^{n+\frac{1}{2}} + \partial_t U^{n-\frac{1}{2}}) + (A^{-1}\delta_t Z^n, Z^n) \\ &= 0 \end{aligned} \tag{69}$$

or

$$\frac{1}{2\Delta t}(\| \partial_t U^{n+\frac{1}{2}} \|^2 - \| \partial_t U^{n-\frac{1}{2}} \|^2) + (A^{-1}\delta_t Z^n, Z^n) = 0. \tag{70}$$

We now wish to estimate $(A^{-1}\delta_t Z^n, Z^n)$. With $Z^{n+\frac{1}{2}} = \frac{Z^n+Z^{n+1}}{2}$ we have

$$Z^{n+1} - Z^{n-1} = 2(Z^{n+\frac{1}{2}} - Z^{n-\frac{1}{2}}),$$

$$Z^n = \frac{(Z^{n+\frac{1}{2}} + Z^{n-\frac{1}{2}})}{2} - \frac{(\Delta t)^2}{4}\partial_t^2 Z^n.$$

Letting $\| \cdot \|_a^2 = (A^{-1}\cdot, \cdot)$, we see that

$$\begin{aligned} (A^{-1}\delta_t Z^n, Z^n) &= \frac{1}{2\Delta t}[\| Z^{n+\frac{1}{2}} \|_a^2 - \| Z^{n-\frac{1}{2}} \|_a^2 \\ &\quad -\frac{(\Delta t)^2}{4}(A^{-1}(\partial_t Z^{n+\frac{1}{2}} + \partial_t Z^{n-\frac{1}{2}}), \partial_t Z^{n+\frac{1}{2}} - \partial_t Z^{n-\frac{1}{2}})] \\ &= \frac{1}{2\Delta t}\{\| Z^{n+\frac{1}{2}} \|_a^2 - \| Z^{n-\frac{1}{2}} \|_a^2 \\ &\quad -\frac{(\Delta t)^2}{4}(\| \partial_t Z^{n+\frac{1}{2}} \|_a^2 - \| \partial_t Z^{n-\frac{1}{2}} \|_a^2)\}. \end{aligned} \tag{71}$$

Combining (70) and (71) and summing on n, n = 1,...,N, we deduce that

$$(72)\qquad \| \partial_t U^{N+\frac{1}{2}} \|^2 + \| Z^{N+\frac{1}{2}} \|_a^2 - \frac{(\Delta t)^2}{4} \| \partial_t Z^{N+\frac{1}{2}} \|_a^2$$
$$= \| \partial_t U^{\frac{1}{2}} \|^2 + \| Z^{\frac{1}{2}} \|_a^2 - \frac{(\Delta t)^2}{4} \| \partial_t Z^{\frac{1}{2}} \|_a^2 .$$

From (63) and the inverse property we have

$$(73)\qquad \| \partial_t Z^{N+\frac{1}{2}} \|_a^2 \leq \frac{C_0 \beta}{h} \| \partial_t U^{N+\frac{1}{2}} \| \| \partial_t Z^{N+\frac{1}{2}} \|_a .$$

Thus,

$$(74)\qquad \frac{(\Delta t)^2}{4} \| \partial_t Z^{N+\frac{1}{2}} \|_a^2 \leq \frac{C_0^2 \beta^2}{4} \left(\frac{\Delta t}{h}\right)^2 \| \partial_t U^{N+\frac{1}{2}} \|^2 .$$

Substituting (74) into (72) we obtain the following stability result.

Theorem 2 *The explicit discrete scheme defined by (34)-(37) with $s = 0$ is stable if $C_0 \Delta t \, \beta < 2h$. That is*

$$\left(1 - (\frac{C_0 \Delta t \, \beta}{2h})^2\right) \| \partial_t U^{N+\frac{1}{2}} \|^2 + \| Z^{N+\frac{1}{2}} \|_a^2$$

is bounded by initial data.

Stability for the general s-discrete scheme for $s \in [0,1]$ will be demonstrated in [8]. The methods are conditionally stable for $0 \leq s < 1/4$ and unconditionally stable for $1/4 \leq s$. For the homogeneous differential equations, the time truncation is minimized by taking $s = 1/6$. In [8] convergence arguments similar those to given here for the continuous-time case will be presented for the s-discrete case. More general boundary conditions will also be addressed.

5 Numerical Experiments

We begin by defining the Raviart-Thomas spaces for Neumann Boundary problems. Let Ω be the rectangular domain $(0,1) \times (0,1)$ and

$$(75)\qquad \Delta_x : 0 = x_0 < x_1 < \ldots < x_{N_x} = 1,$$
$$(76)\qquad \Delta_y : 0 = y_0 < y_1 < \ldots < y_{N_y} = 1,$$

be partitions of $[0,1]$. For such a partition Δ, define

$$M_q^r(\Delta) = \{v \in C^q([0,1]) : v \text{ is a polynomial of degree } \leq r \text{ on each subinterval of } \Delta\}. \tag{77}$$

When $q = -1$ this is taken to be a space of discontinuous piecewise polynomial functions. The Raviart-Thomas spaces are defined as follows:

$$W^{q,r} = M_q^r(\Delta_x) \otimes M_q^r(\Delta_y), \tag{78}$$

$$\tilde{V}^{q,r} = [M_{q+1}^{r+1}(\Delta_x) \otimes M_q^r(\Delta_y)] \times [M_q^r(\Delta_x) \otimes M_{q+1}^{r+1}(\Delta_y)], \tag{79}$$

$$V^{q,r} = \tilde{V}^{q,r} \cap \{v : v \cdot \nu = 0 \text{ on } \partial\Omega\}. \tag{80}$$

Raviart and Thomas [16] show that the above spaces satisfy the inf-sup condition required by [2], [4], etc. For the purpose of our numerical experiments we will restrict our attention to the Raviart-Thomas space of next to the lowest order ($q = -1, r = 1$) and solve the fully discrete problem described in Section 2 with Neumann boundary conditions. For discrete initial conditions we use the L^2 projections of $u(0)$ and $u_t(0)$ into $W^{-1,1}$ defined by (34)-(35). The quantity U^{-1}, needed in the implicit schemes, is defined by the relation $U_t^0 = \frac{U^1 - U^{-1}}{2\Delta t}$ in order to preserve the quadratic convergence in time. Z^{-1} is similarly defined.

Two experiments were conducted to compare the performance of two interesting members of the class of s-discrete schemes given by (34)-(37). Taking $s = 0$ yields a conditionally stable, computationally simple explicit scheme that is second order accurate in time. Setting $s = \frac{1}{4}$ yields an unconditionally stable scheme that is also second order accurate in time. In each case we use constant mesh spacings. Specifically we take $x_{i+1} - x_i$ to be constant at h_x and $y_{i+1} - y_i$ to be constant at h_y, and we take $h_x = h_y = h$.

The first experiment was designed to see if the schemes accurately described the propagation of a radial pulse. We look specifically at the rate of propagation, the height of the crest of the pulse, and a norm of the error computed over the entire spatial domain. For initial conditions we take u_0 a member of the family

$$u_0^\epsilon(x,y) = \begin{cases} P^\epsilon(x^2+y^2) & \text{if } x^2+y^2 \leq \epsilon, \\ 0 & \text{otherwise,} \end{cases} \tag{81}$$

where

$$P^\epsilon(R) = \frac{2}{\epsilon^3}(R - \frac{\epsilon}{2})(R+\epsilon)^2. \tag{82}$$

The parameter ϵ controls the support and sharpness of the pulse. In what follows, we look at two members: $\epsilon = 0.0025$, which we shall refer to as the **sharp pulse,** and $\epsilon = 0.025$, which we shall refer to as the **smooth pulse.** Our first test problem (TP-1) is

$$\begin{aligned} &(83) \qquad & u_{tt} - \Delta u &= 0 \quad \text{in } \Omega \times (0,\infty), \\ &(84) \qquad & \frac{\partial u}{\partial n} &= 0 \quad \text{on } \partial\Omega \times (0,\infty), \\ &(85) \qquad & u(x,y,0) &= u_0^\epsilon(x,y), \\ &(86) \qquad & u_t(x,y,0) &= 0. \end{aligned}$$

Using classical methods we can write the solution $u(x,y,t)$ of equations (83) - (86) as the time derivative of a double integral. For our analysis, we evaluated $u(x,y,t)$ by a combination of analytic techniques and the adaptive quadrature routines of QUADPACK. We shall refer to this calculation of $u(x,y,t)$ as the **comparison solution.**

The mixed finite element procedure (34)-(37) with $W_h = W^{-1,1}$ and $V_h = V^{-1,1}$ was applied to equations (83) - (86) using a variety of spatial grids and time steps. The following discrete L^2 norm evaluated at the Gauss points was used to calculate the error between the comparison solution and the calculated solution.

$$(87) \qquad \|v\|^2 = (r+1)^{-2} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \sum_{l=1}^{r+1} \sum_{k=1}^{r+1} f(x_{il}, y_{jk})^2 h_x^i h_y^j,$$

where x_{il} and y_{jk} are the $r+1$ Gauss points on $[x_{i-1}, x_i]$ and $[y_{j-1}, y_j]$, respectively.

Figures 1 and 2 and Tables 1 through 3 summarize our results. The plots in the figures come from experiments on a 50x50 grid with the smooth pulse initial condition. We see in Figure 1 that for small Δt the wave fronts are nearly circular and coincide well with the comparison solution. As Δt increases, the character of approximation solutions to the pulse change. The pulse becomes thicker and the leading edge looses definition. Also, for $\frac{\Delta t}{h} \geq 1$, we begin to see secondary pulses. While in two spatial dimensions we would expect trailing waves in accord with Huygen's principle, the ones in Figure 1-d and Figure 1-e are approximation anomalies occurring at the wrong time and with the wrong amplitude. Because the fronts are nearly circular, we conclude that the fronts propagate at nearly uniform speed in all directions independent of grid effects.

The maximum height of the pulse over the entire domain is plotted versus time in Figure 2. Given our geometry and initial conditions, three events occur that effect the maximum of $u(\cdot,\cdot,t)$. Initially, the pulse begins to diminish as it spreads out through the domain. Just before $t = 1$, the leading edge of the pulse contacts the boundary and is reflected back. Thus, at points near the boundary there is a contribution to u by that part of the pulse that has already been reflected and that part which is about to be reflected. As a result, the maximum height increases around $t = 1$. The third event occurs near $t = 1.3$ when parts of the pulse reflected off of adjacent boundaries cross each other. From Figure 2, we see that realizations with larger time steps seem to under estimate the height of the wave as it contacts the boundary. This is due in part to the larger time steps smearing out the wave. In doing so, the crest of the wave propagates slower than the leading edge, causing the time of impact and the time of crossing to lag those of the comparison solution. We note that only for $\frac{\Delta t}{h}$ near the stability limit for the explicit method are these phenomena accurately modeled.

Two interesting items deserve particular attention in the tabulation of the errors. First we see that for the sharp pulse ($\epsilon = 0.0025$), the spatial discretization error obscures the quadratic convergence in time expected in the implicit method. On the 50x50 grid the support of the sharp pulse is approximately 8 grid blocks. Because the pulse is so poorly resolved initially, the spatial error dominates the temporal discretization error. For the smoother pulse ($\epsilon = 0.025$), the initial condition is much better resolved, and we see the quadratic in time convergence of the implicit scheme. We note that the error in the implicit scheme is only marginally better than the explicit error with a comparable time step.

Secondly, we verify the spatial super-convergence that the Raviart-Thomas spaces possess at the Gauss points (see [15]). For the experiments depicted

Scheme	$\frac{\Delta t}{h}$	25x25 grid	50x50 grid	100x100 grid
$s = 0$	$\frac{1}{6}$	36.728	7.067	1.848
$s = \frac{1}{4}$	$\frac{1}{8}$	23.008	4.770	1.503
$s = \frac{1}{4}$	$\frac{1}{4}$	40.005	11.353	3.130
$s = \frac{1}{4}$	$\frac{1}{2}$	50.501	21.746	6.240

Table 1: Discrete L^2 Error ($\times 10^5$) for the Sharp Pulse

Scheme	$\frac{\Delta t}{h}$	25x25 grid	50x50 grid	100x100 grid
$s = 0$	$\frac{1}{6}$	17.772	2.693	0.398
$s = \frac{1}{4}$	$\frac{1}{8}$	7.944	1.294	0.261
$s = \frac{1}{4}$	$\frac{1}{4}$	23.752	5.015	0.925
$s = \frac{1}{4}$	$\frac{1}{2}$	48.218	13.595	2.691

Table 2: Discrete L^2 Error ($\times 10^5$) for the Smooth Pulse

Scheme	Δt	50x50 grid	Δt	80x80 grid
$s = 0$	0.00333	2.693	0.00166	0.616
$s = \frac{1}{4}$	0.00250	1.294	0.00125	0.324
$s = \frac{1}{4}$	0.00500	5.015	0.00250	1.064
$s = \frac{1}{4}$	0.01000	13.595	0.00500	3.344

Table 3: Discrete L^2 Error ($\times 10^5$) for the Smooth Pulse

in Table 3, the time steps differ by a factor of two on the two grids, but the 80x80 grid is a spatial refinement of a factor of $\sqrt[3]{4}$ over the 50x50 grid. Because of the super-convergence of the solution at the Gauss points, the error measured by the discrete L^2 norm (87) is $O(h^3 + (\Delta t)^2)$. Therefore, we expect to see a factor of four improvement in the error.

The second experiment compared the ability for the implicit and explicit schemes to refocus a pulse. The values of u and $-u_t$ of the comparison solution for equations (83) – (86) were computed at time $t = 2$ and used as initial conditions for the second test problem. Specifically, the second test problem (TP-2) is

$$\begin{aligned}
&(88) \qquad & v_{tt} - \Delta v &= 0 \quad \text{in } \Omega \times [0, \infty), \\
&(89) \qquad & \frac{\partial v}{\partial n} &= 0 \quad \text{on } \partial\Omega \times [0, \infty), \\
&(90) \qquad & v_0(x, y) &= u(x, y, 2), \\
&(91) \qquad & v_t(x, y) &= -u_t(x, y, 2).
\end{aligned}$$

At $t = 2$, v is u_0^ϵ defined in (81). Figures 3–4 and Tables 4–5 summarize our results.

Figures 3 and 4 depict the cross section of v along the x-axis at time

Scheme	$\frac{\Delta t}{h}$	50x50 grid	100x100 grid
$s = 0$	$\frac{1}{6}$	6.674	1.779
$s = \frac{1}{4}$	$\frac{1}{8}$	5.769	1.333
$s = \frac{1}{4}$	$\frac{1}{4}$	10.965	3.634
$s = \frac{1}{4}$	$\frac{1}{2}$	26.093	6.492

Table 4: Discrete L^2 Error $(\times 10^5)$ for the Sharp Pulse

Scheme	$\frac{\Delta t}{h}$	50x50 grid	100x100 grid
$s = 0$	$\frac{1}{6}$	2.040	0.279
$s = \frac{1}{4}$	$\frac{1}{8}$	1.024	0.134
$s = \frac{1}{4}$	$\frac{1}{4}$	3.997	0.679
$s = \frac{1}{4}$	$\frac{1}{2}$	13.516	2.269

Table 5: Discrete L^2 Error $(\times 10^5)$ for the Smooth Pulse

$t = 2$ for the smooth pulse. In these figures, we observe overshoot in the approximation of the front for larger time steps. Moreover, we see that the support of v is also poorly modeled in the case $\frac{\Delta t}{h} = \frac{1}{2}$. From the tabulation of the errors, we again see that while the implicit is unconditionally stable, we must severly limit the time step to reproduce the same error as the explicit scheme. As in the first test problem, we also see that spatial error dominates the quadratic convergence in time for the sharper pulse.

Limits on time step size are imposed by two considerations, stability and accuracy. We have seen for the next to the lowest order mixed method that while the implicit method with $s = \frac{1}{4}$ possesses no stability limitation on the time step, accuracy considerations impose an equally severe restriction on the time step. While more test problems need to be considered, our results seem to suggest that for this higher-order s-discrete method, the computationally cheaper explicit method combined with some *a priori* knowledge concerning stability is preferable to the implicit method.

References

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, 1975.

[2] D. Arnold, *Mixed finite element methods for elliptic problems*, this Volume.

[3] D. Arnold and F. Brezzi, *Mixed and nonconforminy finite element methods: implementation, postprocessing and error estimates*, Math. Model. and Numer. Anal., 19 (1985), pp. 7–32.

[4] I. Babuska and J. E. Osborn, *Generalized finite element methods: their performance and their relation to mixed methods*, SIAM J. Numer. Anal., 20 (1983), pp. 510–536.

[5] G. A. Baker, *Error Estimates for finite element methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 13 (1976), pp. 564–576.

[6] F. Brezzi, *On the existence, uniqueness, and approximation of saddle point problems arising from Lagrangian multipliers*, RAIRO Anal. Numer., 8-32 (1974), pp. 129–151.

[7] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math. 47 (1985), pp. 217–235.

[8] L. Cowsar, T. Dupont, M. F. Wheeler, *Mixed finite element methods for second order hyperbolic problems, in preparation.*

[9] J. Douglas, Jr., R. Ewing, and M. F. Wheeler, *The approximation of the pressure by a mixed method in the simulation of miscible displacement*, RAIRO Anal. Numer., 17 (1983), pp. 17–33.

[10] T. Dupont, R. Glowinski, W. Kinton, M. F. Wheeler, *Mixed finite element methods in flow problems*, Hunstville, Alabama, April 1989.

[11] T. Dupont, *L^2-estimates for Galerkin Methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 10 (1973), pp. 880–889.

[12] R. E. Ewing, T. F. Russell, and M. F. Wheeler, *Convergence analysis of an approximation of miscible displacement in porous media by*

mixed finite elements and a modified method of characteristics, Computer Methods in Applied Mechanics and Engineering 47, pp. 73–92 (1984).

[13] T. Geveci, *On the application of mixed finite element methods to the wave equation*, Math. Model. and Numer. Ana., 22 (1988), pp. 243–250.

[14] J. T. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, 1972.

[15] M. Nakata, A. Weiser, and M. F. Wheeler, *Some superconvergence results for mixed finite element methods for elliptic problems on rectangular domains*, The Mathematics of Finite Elements and Applications (MAFELAP–1984), (Ed. J. R. Whiteman), Academic Press, London 1984.

[16] P. A. Raviart and J. M. Thomas, *A mixed finite element method for* 2^{nd} *order elliptic problems,* in *Mathematical Aspects of the Finite Element Method*, Springer Lecture Notes in Mathematics 606 (1977), Springer-Verlag, Heidelberg.

[17] E. Stein, *Singular Integrals and Differcntiability Properties of Functions*, Princeton University Press, 1970.

CAPTIONS FOR FIGURES

Figure 1a : Comparison Solution

Figure 1b : $s = 0 \quad \frac{\Delta t}{h} = \frac{1}{6}$

Figure 1c : $s = \frac{1}{4} \quad \frac{\Delta t}{h} = \frac{1}{8}$

Figure 1d : $s = \frac{1}{4} \quad \frac{\Delta t}{h} = \frac{1}{4}$

Figure 1e : $s = \frac{1}{4} \quad \frac{\Delta t}{h} = \frac{1}{2}$

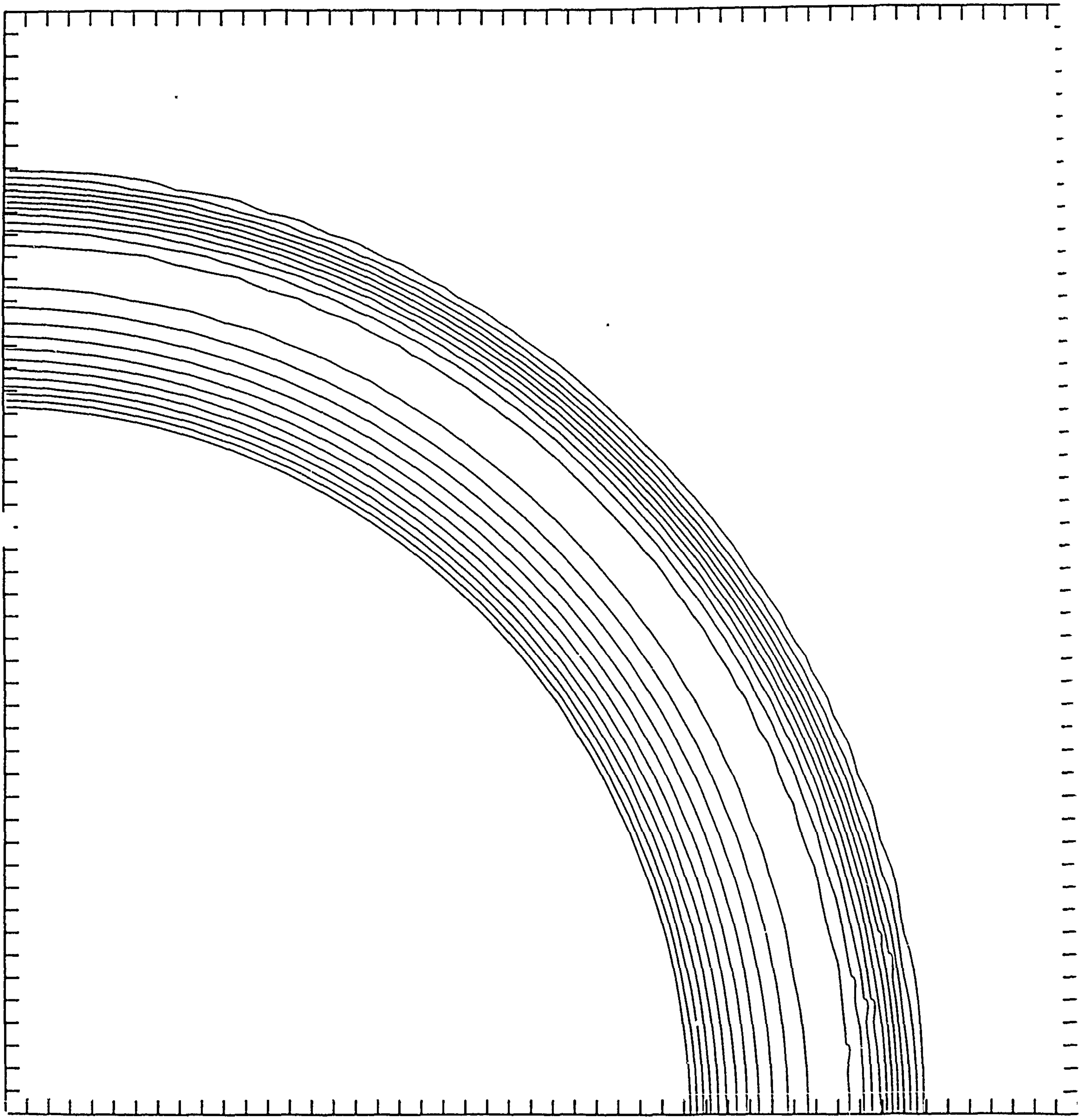
Figure 1f : $s = \frac{1}{4} \quad \frac{\Delta t}{h} = 1$

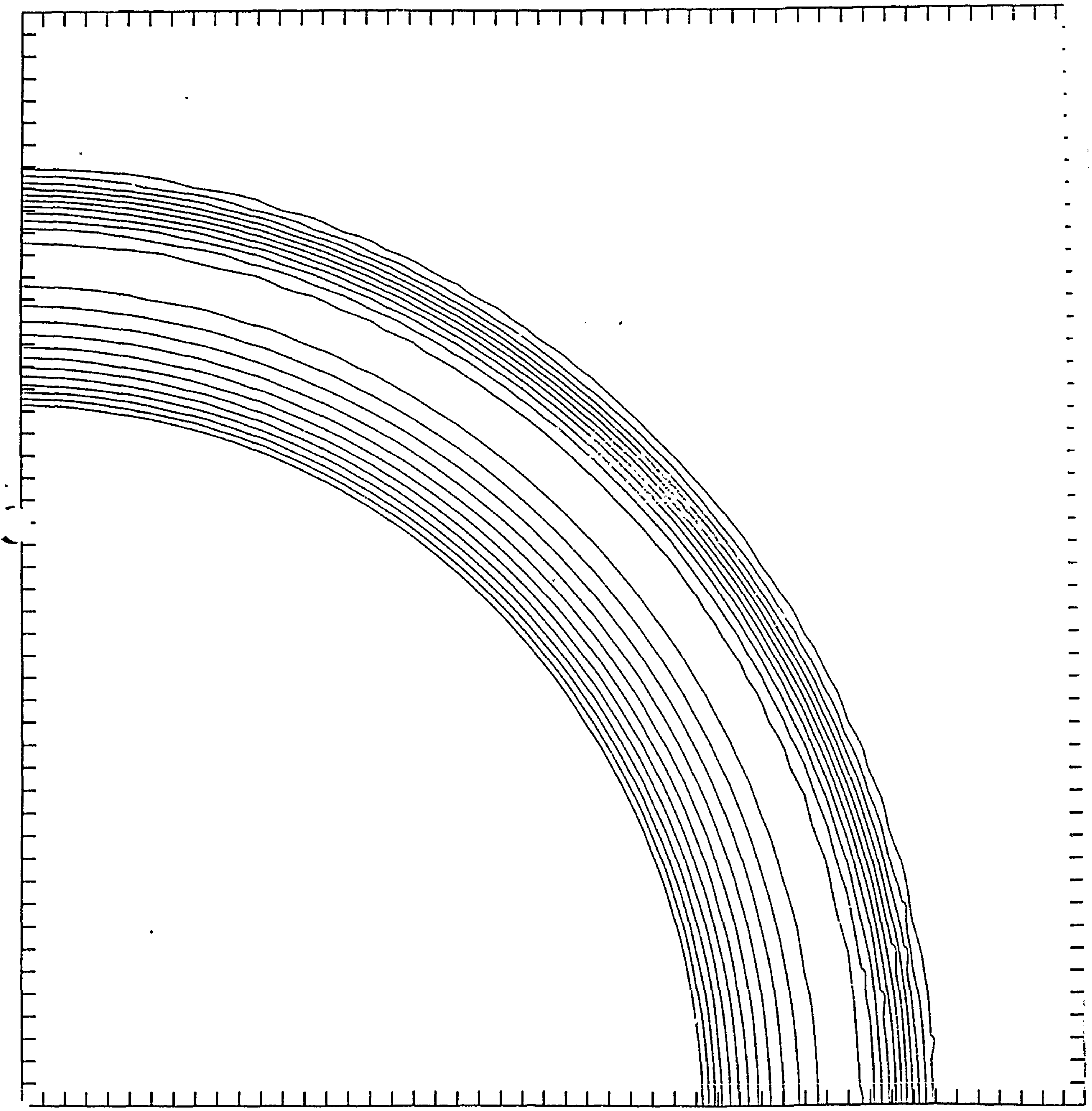
Figure 1g : $s = \frac{1}{4} \quad \frac{\Delta t}{h} = 2$

Figure 2

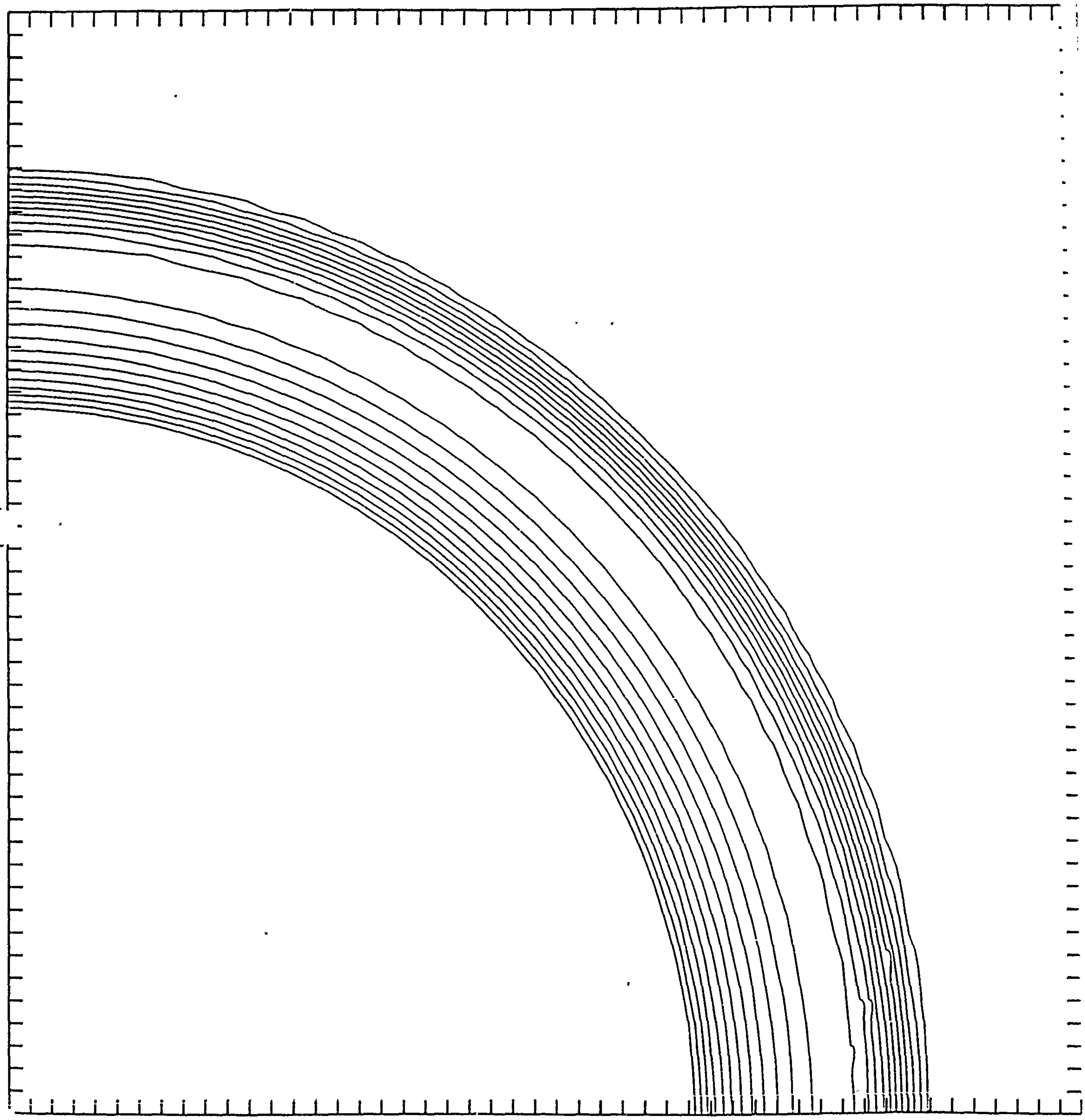
Figure 3

Figure 4





F. [illegible]



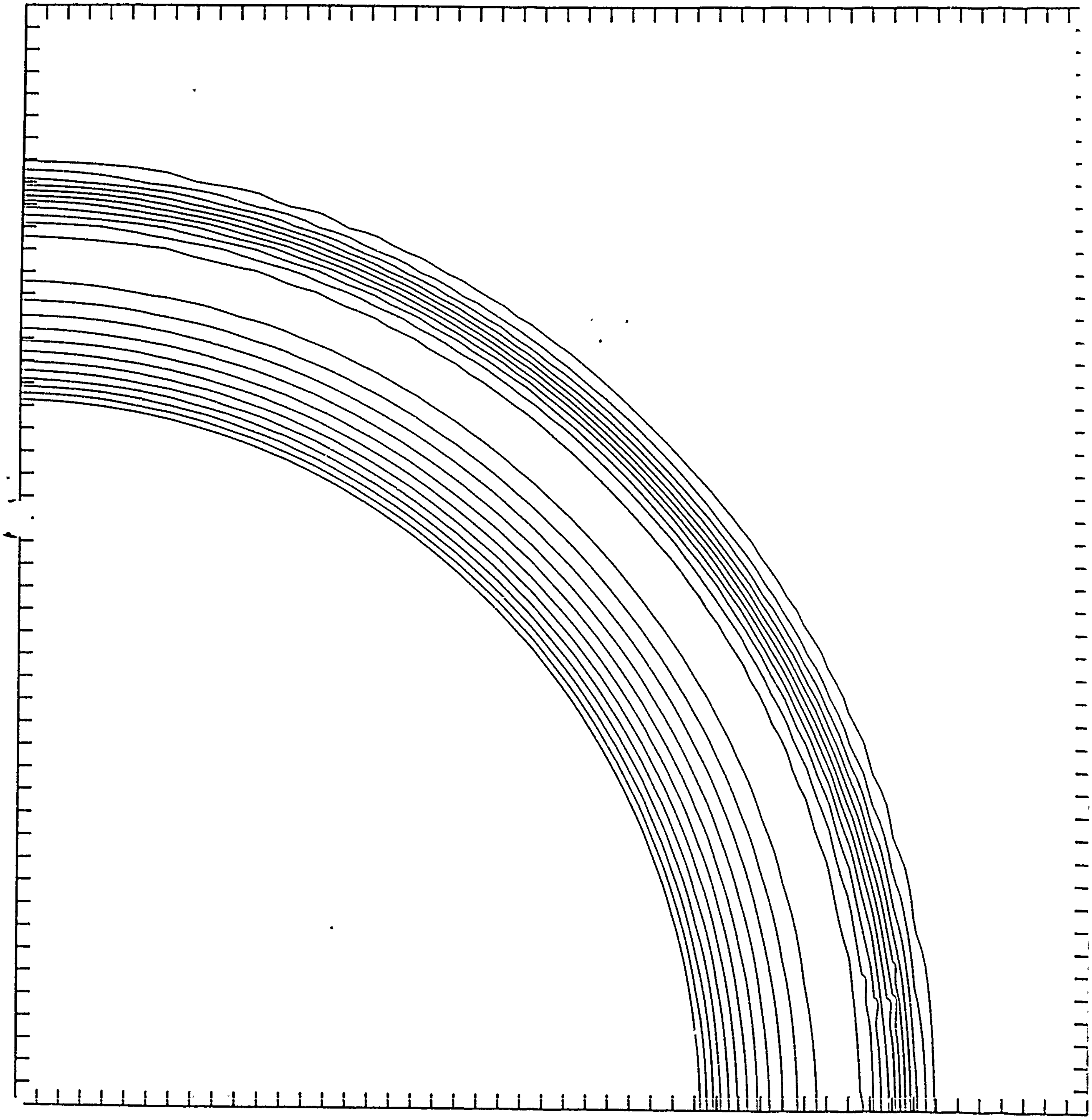
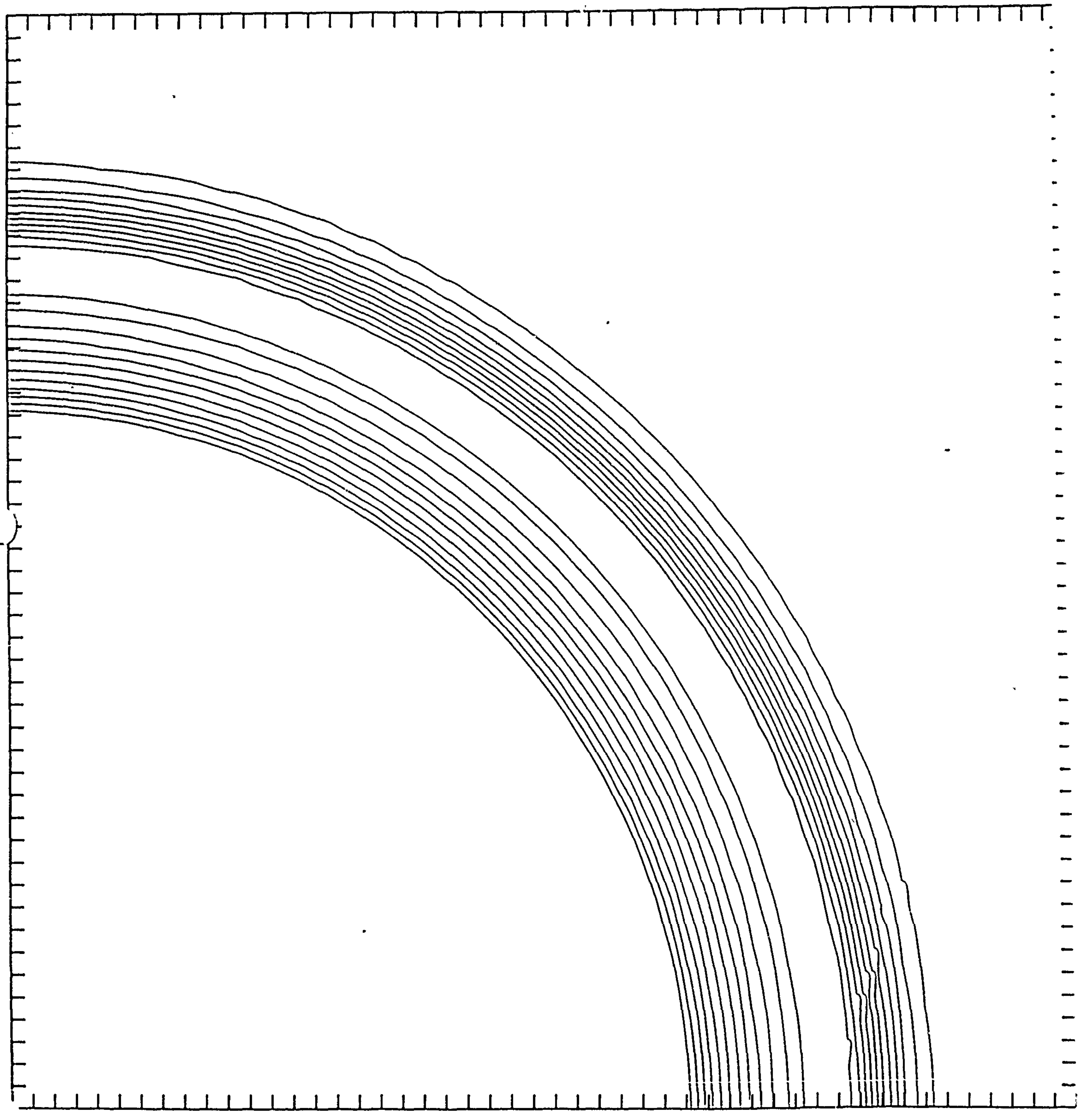
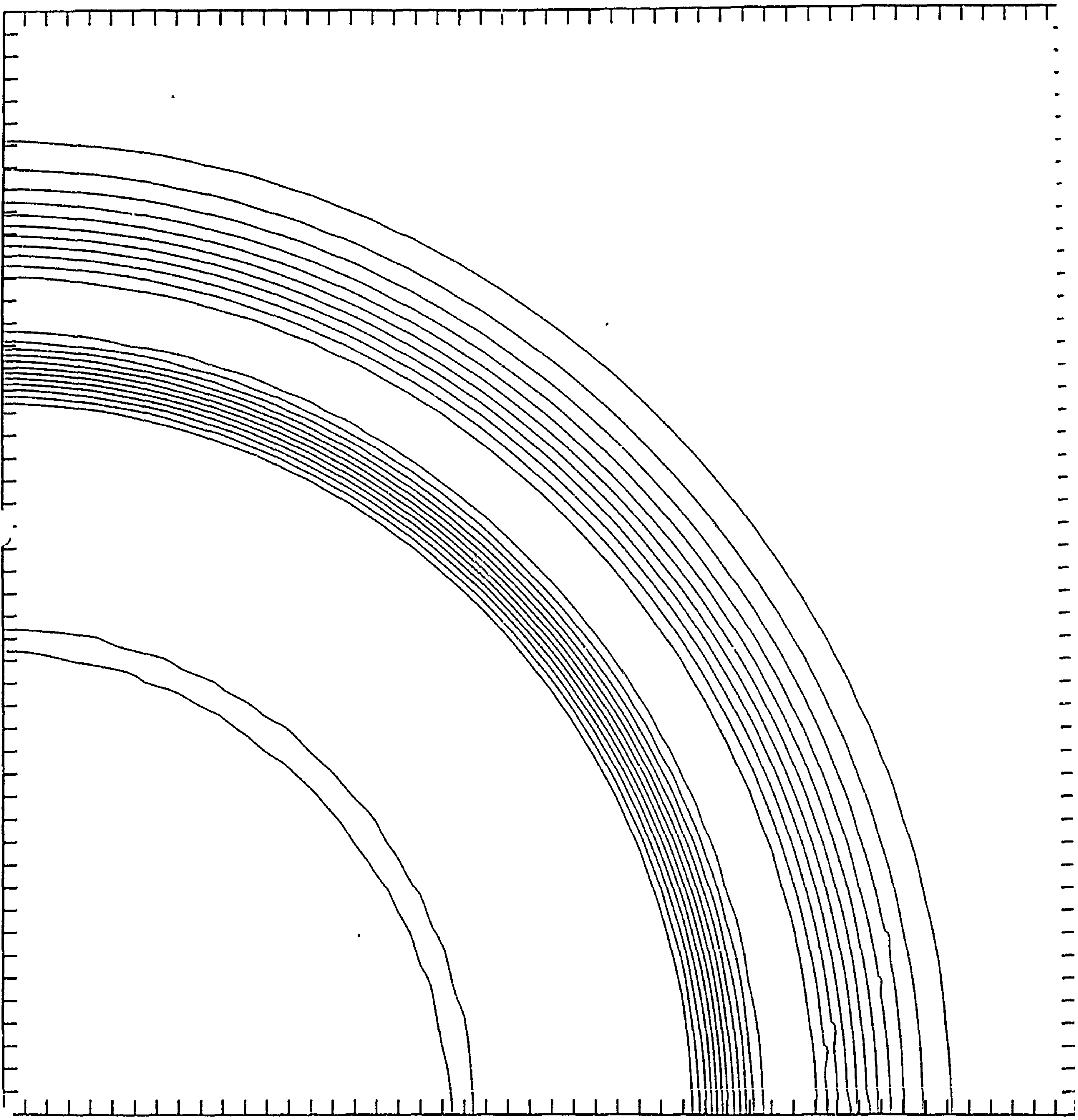
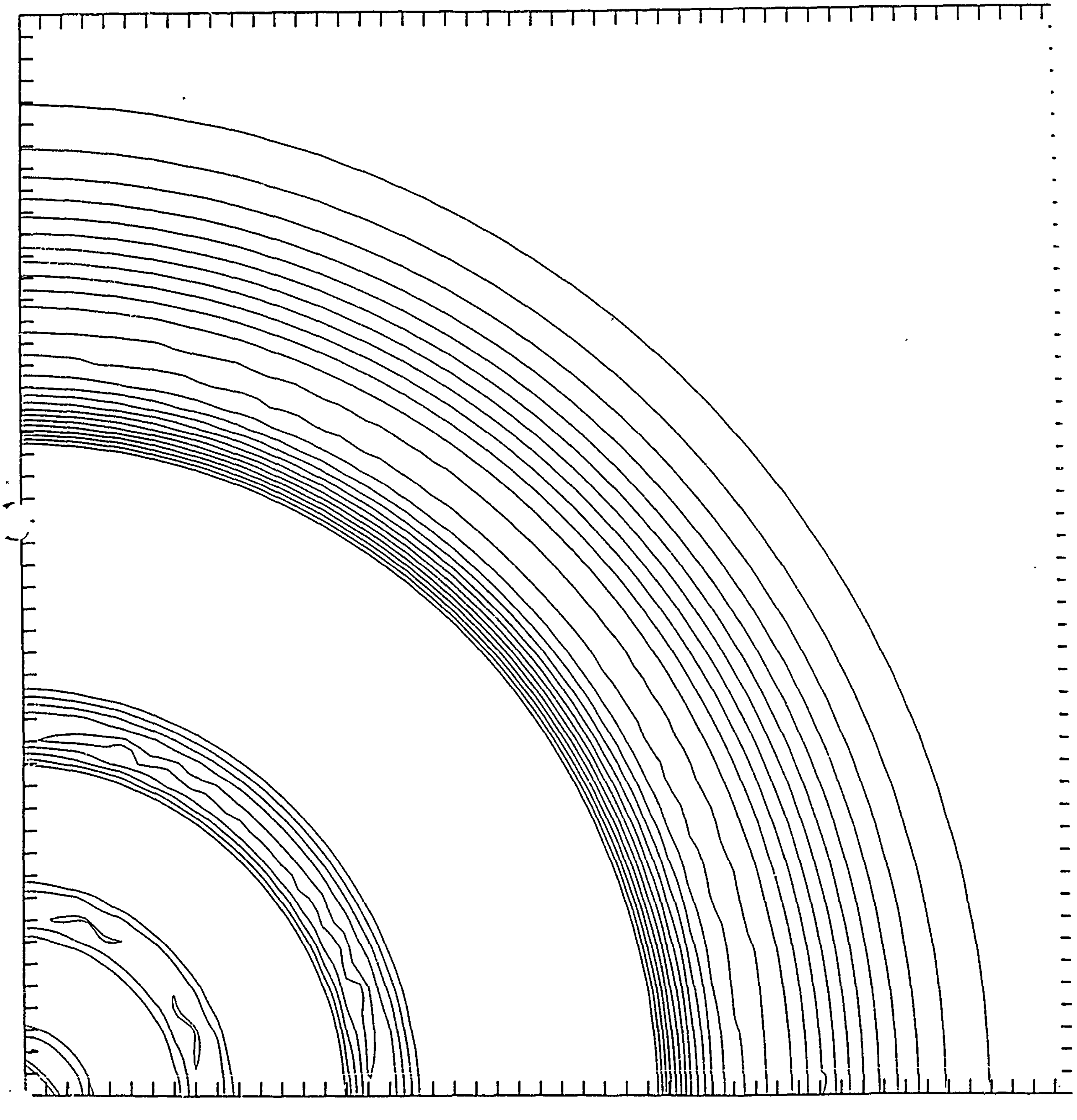
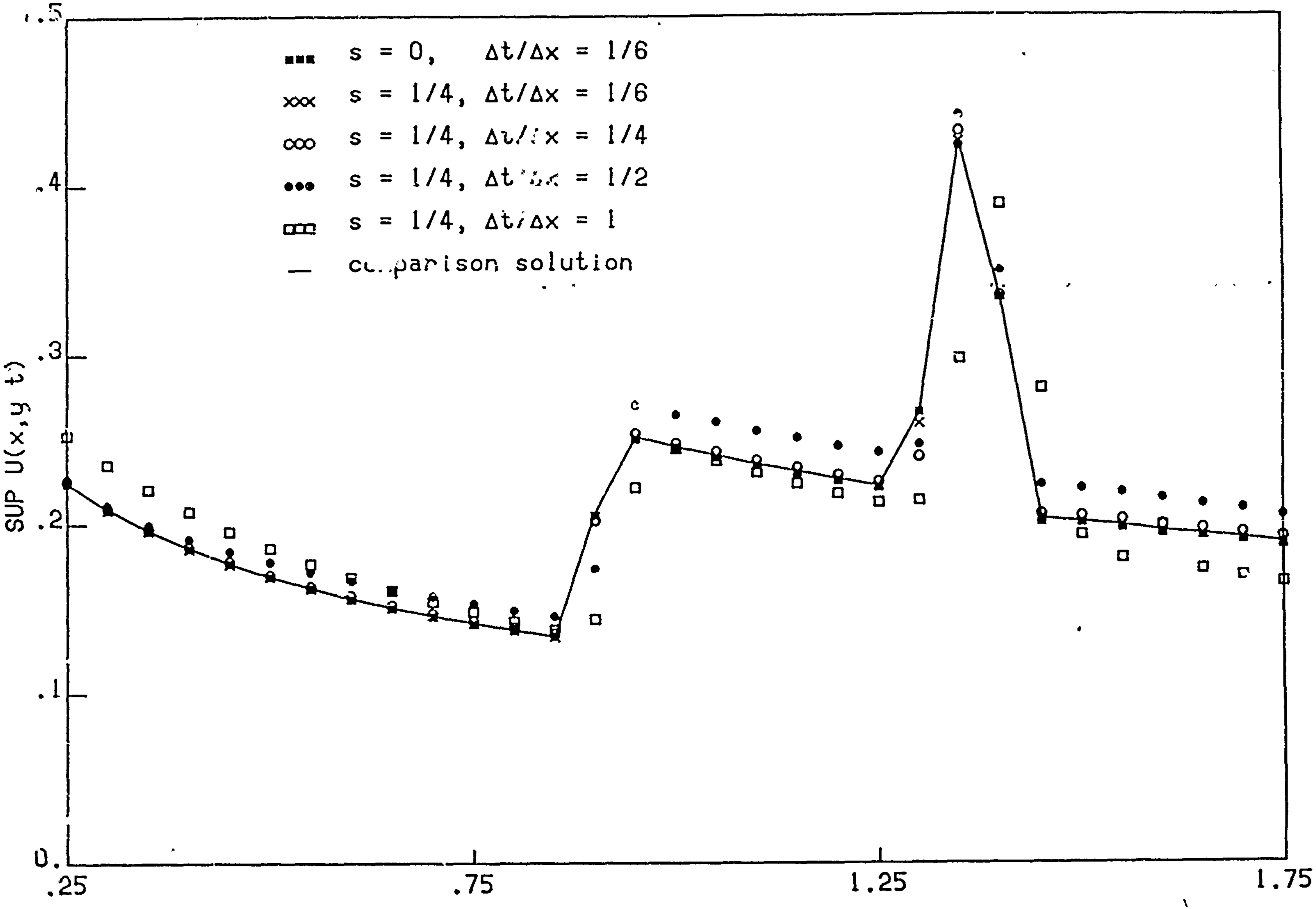


Fig 1-c



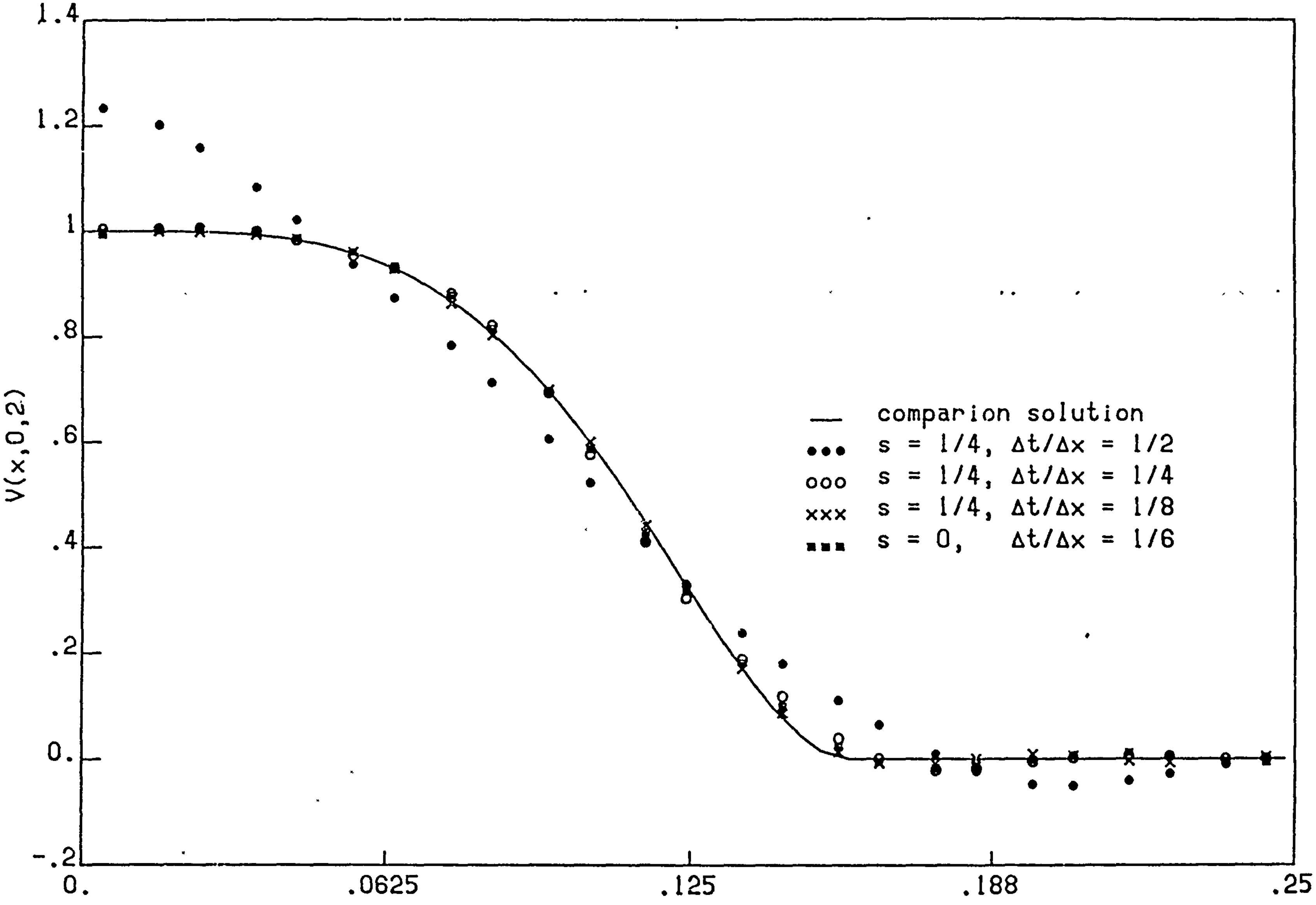






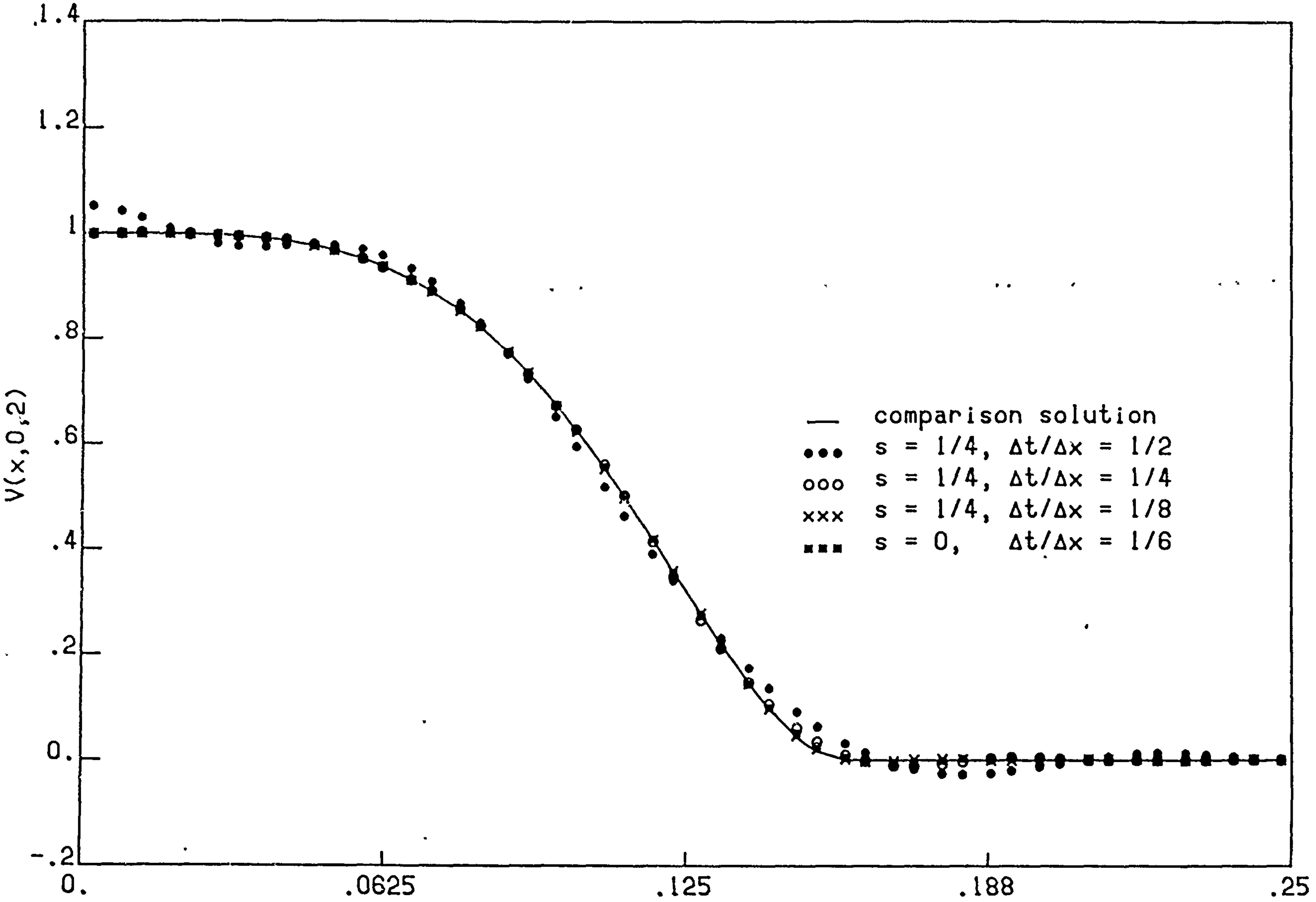
HEIGHT OF WAVE CREST VERSUS TIME, TEST PROBLEM I

Fig. 7.



CROSS SECTION OF SOLUTION, GRID 50 X 50

FIG. 7



CROSS SECTION OF SOLUTION, GRID 100 X 100

Accuracy and Adaptivity in the Numerical Analysis of Thin-Walled Structures

L. Plank E. Stein [1] D. Bischoff [2]

[1] Prof. Dr.-Ing. E. Stein, Dipl. Math. L. Plank, Institut für Baumechanik und Numerische Mechanik, University of Hannover, Appelstraße 9a, D-3000 Hannover 1

[2] PD Dr.Ing.-habil, Dr.rer.nat. D.Bischoff, Control Data Corporation, Tiergartenstraae 95, D-3000 Hannover 1

Abstract

An integrated concept for more accurate robust and reliable FE-computations of thin-walled metal structures with elasto-plastic large deformations subject to static loads is presented. It includes adaptive mesh refinement using a-priori and a-posteriori criteria and the use of a multigrid solver. We have developed a mesh refinement procedure that allows the combination of mesh adaptation and multigrid methods for arbitrarily complex structures. Illustrated examples for for a-priori and a-posteriori mesh refinements are included.

1. Introduction

The topic of this paper is the development of a concept for an efficient and reliable static FE-analysis of large thin-walled structures as they appear in industrial applications. In this context the control of the FE-mesh using a-priori and a-posterioro criteria is necessary in order to obtain optimal convergence of the approximate solution. The structural elements are folded plates or shells consisting of metals, especially steel, with **stiffening** plates or both sides. Besides linear-elastic inembrane and bending stresses there are effects of large displacements and relations of the structural elements and the influence of plastic deformations on the stress states. The formulations of all these mechanical properties and incertainties is given in chapter 2 followed by the finite element discretization methods in chapter 3. Quadrilateral and triangular isoparametric elements are used where only regular nodes are permitted in the mesh refining process.

A-priori and a-posteriori refinement criteria are developed in chapter 4. The a-priori criteria control the shape of the elements and the deviations from the geometry. The a-posteriori criteria for h-adaptivity contain reliable error estimators for linear problems and heuristic criteria for geometrical nonlinearity and plastifying zones as well.

A careful choice of the solution algorithms for large systems of equations is another important point, as the computational effect of standard equations solves grows superlinearly with the number of degrees of freedom. In the industrial use of the finite-element method the practical applicability of new methods is even more important than their theoretical advantages. The generation of a finite element mesh from CAD data using standard mesh generators is a time-consuming task and the additional effect of adaptive refinement or generation of a sequence of meshes as well. Chapter 5 gives an overview of equation solvers with special emphasis on multigrid methods for FE-computations.

In chapter 6 the mesh refinement technique is discussed using a newly developed "Shape Preserving Recursive Mapped Meshing".

Summarizing an integrated [illegible]on refining concept using a-priori and a-posteriori indicators is presented, and details of the refinement and solution techniques are described. A couple of illustrative examples with computer plots are included within the corresponding chapters which demonstrate the efficiency of the whole concept.

which results in the symmetry of $\mathbf{T}_B$

$$\mathbf{T}_B = (\mathbf{R}^T)\mathbf{T}(\mathbf{R}) = \mathbf{T}_B^T = \mathbf{S} \qquad (7)$$

where S denotes the 2nd Piola–Kirchhoff stress tensor. The principle of virtual work now reads

$$\int_V (\mathbf{T}_B \cdot \delta\mathbf{U} - \rho_0 \mathbf{b} \cdot \eta) dV - \int_{A_\sigma} \bar{t}_0 \cdot \eta dA_\sigma \qquad (8)$$

The stress resultants are defined by

$$\mathbf{n}_\alpha := \int_{-\frac{h}{2}}^{\frac{h}{2}} \mathbf{T}_\alpha d\xi \quad , \quad \mathbf{m}_\alpha := \int_{-\frac{h}{2}}^{\frac{h}{2}} (\mathbf{x} - \mathbf{x}_0) \times \mathbf{T}_\alpha d\xi \qquad (9)$$

The principle of virtual work can be formulated in the initial configuration

$$\delta\mathbf{W} = \int_A (\mathbf{N}_\alpha \cdot \delta\mathbf{\Gamma}_\alpha + \mathbf{M}_\alpha \cdot \delta\mathbf{K}_\alpha) dA \qquad (10)$$

using the transformations

$$\mathbf{N}_\alpha = \mathbf{R}^T \mathbf{n}_\alpha \ , \ \mathbf{M}_\alpha = \mathbf{R}^T \mathbf{m}_\alpha. \qquad (11)$$

for the stress resultants and

$$\begin{aligned} \mathbf{\Gamma}_\alpha(S_\alpha) &= \mathbf{R}^T \mathbf{x}_{0,\alpha} - \mathbf{t}_\alpha && (12) \\ \mathbf{K}_\alpha(S_\alpha) &= \mathbf{R}^T \omega_\alpha && (13) \end{aligned}$$

for the strains.

We prefer to work in the corotational configuration, which is defined by a decomposition (or operator split) of the motion into a rigid body motion and a deformation

$$\chi = \chi^D \circ \chi^C \qquad (14)$$

Fig. 2 : 2D-draft of the rigid body rotation χ^C and the subsequent deformation χ^D of an element

and correspondingly

$$\mathbf{R}^D = \mathbf{R}(\mathbf{R}^C)^T \qquad (15)$$

(The superscript 'D' denotes the deformational part, 'C' the corotational (rigid body) part). The stress resultats now transform as follows:

$$\mathbf{N}_\alpha^C = (R^D)^T \mathbf{n}_\alpha \quad \mathbf{M}_\alpha^C = (R^D)^T \mathbf{m}_\alpha \quad . \qquad (16)$$

For the evaluation of strains, the rotation tensor

$$\mathbf{R} = e^{\Psi} \tag{17}$$

is linearized

$$\mathbf{R}^D = \mathbf{1} + \Psi \; ; \; (\mathbf{R}^D)^T = \mathbf{1} - \Psi \quad . \tag{18}$$

This linearization is only admissible, if Ψ, which describes the element distortion, is small. This assumption must be controlled during the computation. This tensor is used in chapter 4 as the basis of an error indicator detecting the geometrical nonlinearity of the FE-solution. An extented treatment of the geomtrically nonlinear kinematics is given in [3] (see also [4]).

2.3 Material laws

As we restrict ourselves to small strains, we apply Hooke's law in the corotational formulation for elastic deformations.

Inelastic behaviour is described by an elasto-plastic material law with a von Mises yield condition and isotropic or kinematical hardening

$$F(\tau, \alpha, e_v) = (\tau - \alpha)^D \cdot (\tau - \alpha)^D - \frac{2}{3}\overline{\sigma}^2(e_v) \leq 0 \tag{19}$$

$$J_2 = \frac{1}{2}(\tau - \alpha)^2(\tau - \alpha) \tag{20}$$

and an associated flow rule:

$$\dot{\varepsilon}^{pl} = \dot{\lambda}(\tau - \alpha) \quad . \tag{21}$$

The material laws only approximate the physical situation and leave to be verified by experiments. The elasto-plastic material law is time dependent, which complicates the mathematical analysis. The case of ideal plasticity (without hardening) is especially critical because slip lines can occur. J_2 as an invariant of the stress state is used in chapter 4 as an indicator for mesh refinement with respect to plastic deformation.

3 Discretization

The finite element method is by now well established. Nevertheless, even for linear boundary value problems (plate bending) there is still a need for reliable elements with proven accuracy. Virtually no error analysis is available for nonlinear FE formulations for thin-walled structures.

3.1 Finite element formulations

3.1.1 Linear element formulations

In the linear theory, elements for folded plate structures can be treated as a linear combination of a membrane and a plate bending element. For practical reasons, an artificial stiffness is introduced for the normal rotations. We normally use four-node elements, but — in the case of adaptive mesh refinement — triangular elements cannot always be avoided. For the membrane part, we use the usual bilinear (or linear) shape functions.

For the bending part, we normally use the Reissner/Mindlin plate equation. Locking is avoided by a special (reduced) interpolation for shear strains (Bathe/Dvorkin [5]).

Fig. 3 : Parameter plane of the bilinear element

The shear strains at the midside points are evaluated according to the Reissner-Mindlin plate equations, all other values are obtained by linear interpolation.

$$\begin{aligned} \tilde{\epsilon}_{rz} &= \frac{1}{2}(1-s)\bar{\epsilon}^{A}_{rz} + \frac{1}{2}(1+s)\bar{\epsilon}^{C}_{rz} \\ \tilde{\epsilon}_{sz} &= \frac{1}{2}(1-r)\bar{\epsilon}^{D}_{sz} + \frac{1}{2}(1+r)\bar{\epsilon}^{B}_{sz} \end{aligned} \qquad (22)$$

An a-priori convergence estimate was given by Bathe and Brezzi [6]. It is only valid for rectangular elements, and the estimate for the shear strains depends on the plate thickness. Numerical experiments show low accuracy of the shear strains if the element mesh is distorted, i. e. if adjacent elements have different shapes.

The popular plate elements with selectively reduced integration fit into this context; for bilinear shape functions, the reduced integration is equivalent to constant interpolation of shear strains. It's most serious shortcoming is the presence of zero-energy modes.

There are promising developments of new elements for the Reissner-Mindlin plate equation; we mention an triangular element with optimal error estimates by Arnold and

Falk [7] and mixed elements based on the work of Arnold, Brezzi and Douglas [8] (see also Stein,Rolfes [9]).

A-priori convergence estimates of FE-methods for quasi-uniform meshes show, that the rate of convergence depends on the approximation properties of the shape functions and the smoothness of the solution. In practical applications, the approximated solution often contains singularities (caused by nonsmooth boundaries, loads and the change of boundary conditions), and the performance of FE-methods deteriorates. It can be shown, however, that the optimal rate of convergence can be restored, if adapted non-uniform meshes are used (see I. Babuška, R. B. Kellogg, and J. Pitkäranta [10]).

3.1.2 Nonlinear Finite Element Formulations

A detailed description of the finite element formulation is given in Lambertz [4]. Here, we describe only the treatment of the rotation and curvature of the elements. The orthogonal matrices in the element nodes are calculated from three rotational degrees of freedom $\psi_{1k}, \psi_{2k}, \psi_{3k}$ by an exponential map

$$\mathbf{R} = e^{\Psi} \tag{23}$$

which can be done by Euler parameters (unit quaternions)

$$q_{0k} = \cos\frac{1}{2}\|\psi_k\| \tag{24}$$

$$q_{ik} = \frac{\psi_{ik}}{\|\psi_k\|}\sin\frac{1}{2}\|\psi_k\| \qquad i = 1,2,3 \tag{25}$$

The description of R using these parameters is singularity-free. As the orthogonal rotation tensor is also needed in the interior of the elements, it must be interpolated from the nodal values, which can be done by calculation of the three independent Rodrigues parameters from the Euler parameters, bilinear interpolation of the Rodrigues parameters and calculation of the orthogonal matrices from the Rodrigues parameters. The components of the curvature vector in the integration points can now be computed as follows:

$$\mathbf{\Gamma}_\alpha = \mathbf{R}^T(\mathbf{x}_{0,\alpha} - \mathbf{a}_\alpha) \tag{26}$$

$$\mathbf{K}_\alpha = \mathbf{R}^T\boldsymbol{\omega}_\alpha \tag{27}$$

To avoid shear locking we use selectively reduced integration. At present there is no a-priori error analysis of the nonlinear formulation.

3.2 Plasticity

A detailed mathematical analysis of the elasto-plastic initial-boundary value problem is complicated, because the stresses are bounded in the $\mathcal{L}_\infty$-Norm ($\mathcal{L}_\infty$ is not reflexive). Details can be found in the papers of Strang, Matthies and Temam [11] or Johnson [12]. Without satisfactory a-priori convergence erstimates for the elasto-plastic FEM analysis, an a-posteriori error estimation cannot be given.

Instead, we investigate the additional discretization concepts, that are introduced for the treatment of elasto-plastic behaviour, separately. For the rate constitutive equation, we use an implicit time integration (projection method), which guarantees, that the yield condition is satisfied. Like the well-known implicit Euler-method, this procedure is stable.

(A stability analysis was given by Ortiz and Popov [13]). An accuracy control of the time integration can be done using standard techniques for ordenary differential equations.

In the case of plate bending, the distribution of the stresses over the plate thickness is not uniform. We introduce a layer model, where the material law is integrated in several layer points by the projection method, and the stress resultants are computed from the projected stresses. The numerical quadrature introduces an additional error, that can be easily controlled. For more details, we refer to Stein, Lambertz, and Plank [14], Lambertz [4]. A different approach is an Ilyushin-type material law for stress resultants, that is less accurate, because the error can not be controlled easily and an adaptation is not possible.

4. Criteria for FEM geometry representations

The criteria to determine whether a FE-mesh or an element in a mesh is of good or bad quality, can be subdivided into two categories : a-priori criteria which can be evaluated without knowing the FE-solution and a-posteriori criteria, which allow to estimate the deviation of a local error from an average error.

4.1 A-priori criteria

A priori criteria concern mainly the shape of the elements and the approximation of the geometry of the analysis model. They result to a large extent from the theory of error estimation for approximate solutions of partial differential equations, which are discretized with isoparametric elements [14]. Typically we have a following estimate for the error of the FE-solution $u_h n$.

$$\| u - u_h \| \leq C \cdot h^p \qquad \text{with h: characteristic element length.}$$

This estimate is valid if:

a) the deviation of the elements from the analysis model is relatively small with respect to the element size;

b) the element shape does not differ too much from a square;

c) the elements do not deviate much from the plane;

d) the elements are conforming .

These criteria together with some heuristics can be controlled a-prori.

Deviation from the analysis model :

For piecewise plane analysis models - as they occur frequently - it is sufficient to control the deviation of the element edges from curved boundaries of the analysis model. Element edges which do not satisfy the prescribed quality criteria should be refined until the desired goal is reached. T ·esult of such a check could be seen at a plate with an elliptical hole. If the boundary c ει hole is approximated such that the deviation in the middle of the element edges ar smaller than one percent of the corresponding edge length, the goal is reached after 12 refinement steps. The resulting mesh can be seen in Fig. 4.

Fig. 4: Plate with ...

If free-form surfaces are to be discretized one should check the deviation from the analysis model in every element. An unsatisfactory approximation can be improved by successive refinement of whole elements. The result of such a procedure can be seen at the discretization of the wing of an aircraft.

Fig. 5 : Wing of an aircraft : refined with controlled surface deviation

Deviation from the quadratical square shape

The ideal shape of a quadratical element is the square. There a several waxs to measure the deviation from a square [15], [16]. A hint for the deviation from the quadratical shape is the angle between the two lines which bisect opposite edges. The smaller one is usually called " skewness" and has an amount between 0°and 90°. The deviation of skewness from 90° gives evidence of the unsymmetry of an element (see Fig.: 6))

As indication for sharp and obtuse angles or longish elements, respectively, one can use the ratio of the length of the perpendical line from the midsides of the edges to the diagonals to the length of the whole corresponding diagonal. The minimum of these values is called "aspect ratio" and amounts to 0.25 in the best case (square) (See Fig. 6)). Usually it is between 0. and 0.25.

One gets a suggestion for tapered elements, if the element is subdivided into 4 triangular surfaces by means of the 4 corner points and the midpoint (i.e. the intersection of lines which bisect opposite edges). The ratio of the smallest of these surfaces to the whole surface is called "taper" and has an amount between 0. and 0.25. (The taper of square is 0.25)(See Fig. 6)

Fig. 6 : Element distortion : Skewness, aspect ratio, taper

In these cases no general refinement strategy can be suggested to improve the the element quality. Nevertheless the goal of a strategy should be to intersect longish elements, to isolate sharp angles in triangles and to divide obtuse angles. The simplest possible actions are the division of elements into two or three quadrilaterals respectively, the division to one triangle and one quadrilateral and the division into two triangles and one quadrilateral. The decision which of these strategies should be applied needs a more careful investigation.

Deviation of the planar shape

As a measure for the deviation from the planar quadrilateral shape one uses normally the maximum angle between the element edges and the plane given by the two lines wich bisect opposite edges. It is called "warpage" and is 0 for plane elements. An improvement of the element quality needs, just like in the cases described above, a more careful investigation of the reasons for the warpage. As a refinement strategy all the cases listed above come into question.

Fig. 7 : Element distortion : Warpage

Conformity of elements

Since nonconforming elements produce meaningless results (at least in standard finite element codes), one must be able to detect such nonconforming elements. This is possible by distinguishing between topological and geometrical edges, where the topological edges are defined by the unique neighbourhood. The criteria for conformiug elements is, that every geometrical edge corresponds to exactly one topological edge. In Fig. 8 you see a nonconforming element A, since the geometric edge a corresponds to the topological edges α and β.

Fig. 8 : Geometry ..

Heuristic criteria

Heuristic criteria can be derived by a thorough examination of the analysis model (see e.g.[18]). Especially for linear problems, critical zones can be detected in advance (e.g. singularities in corners of the model; discontinuous material properties, loads and boundary conditions). In the same way, an experienced structural engineer can give meaningfull upper bounds for the maximum diameter of the elements. By specifying such criteria, an examination of a-posteriori criteria can be anticipated. However it must be said, that these heuristic criteria stick very closely to the description of the analysis model and depend on the skill of the user. In [18] this procedure is explicitly labelled "expert-system like" and it should be used with care.

The application of these criteria to a complex structure is demonstrated in the Fig.9 .

Fig. 9 : Element mesh for a car body (Audi)

4.2 A-posteriori Adaptation Criteria

4.2.1 Linear Equations

We consider an elliptic boundary value problem of second order:

$$\mathrm{Lu} = \mathrm{f} \quad \text{in} \quad \Omega(+\text{boundary conditions}) \tag{28}$$

For these type of problems, a theory of reliable error estimators has been developed by Babuska, Rheinboldt et el. (see [19],[20],[21]). In this approach, it is not necessary to evaluate an error function $e^{(h)} = u - u_h$; the energy norm of the error is directly approximated. Estimation of the energy norm of the error by a sum of local (elementwise) error indicators:

$$c_1 \sum_{i=1}^{m} \| \wp_i(e^{(h)}) \|_E^2 \leq \| e^{(h)} \|_E^2 \leq c_2 \sum_{i=1}^{m} \| \wp_i(e^{(h)}) \|_E^2 \tag{29}$$

The indicator $\| \wp_i(e^{(h)}) \|_E^2$ can be expressed as:

$$\begin{aligned} \| \wp_i(e^{(h)}) \|_E^2 &\approx h_i^2 \int_{\Omega_i} (Lu^{(h)} - f)^T (Lu^{(h)} - f) dx \\ &+ h_i \int_{\partial\Omega_i} J(\sigma^{(h)})^T J(\sigma^{(h)}) ds, \end{aligned} \tag{30}$$

where
$Lu^{(h)} - f =$ interior residual (can be neglected, if (bi)linear elements are used)
$J(\sigma^{(h)}) :=$ "jump" of stresses across adjacent elements

This can be directly applied to the membrane equation and to the Reissner/Mindlin plate equation (see Rank [22]). For the four-node element as presented in chapter 3, it suffices to evaluate the jump of the stress resultans

$$(n_1, n_2, n_{12}, m_1, m_2, m_{12}, q_1, q_2) \tag{31}$$

along the element sides and integrate the squares of the differences numerically.

The contributions from the membrane equation $(n_1, n_2, n_{12})$and the plate equation $(m_1, m_2, m_{12}, q_1, q_2)$ must be evaluated differently, which results in error indicators η_M and η_B (for the bending part) and corresponding thresbolds $\overline{\eta}_M$ and $\overline{\eta}_B$. An element is refined, if $\eta_M > \overline{\eta}_M$ or $\eta_B > \overline{\eta}_B$.

The second example is a simply supported rhombic plate under distributed load. The angle is 60°. The solution has a singularity in the obtuse corner : $u \in H^{5/2-\epsilon} > 0$ (Kirchhoff plate theory). This problem was used as a benchmark in the DFG-Schwerpunkt "Nichtlineare Berechnungen im Konstruktiven Ingenieuerbau" (see [23]). For this computation the Reissner/Mindlin plate bending element by Bathe/Dvorkin [5] was used. The reference solution was taken from the above mentioned benchmark (contribution Rannacher [23], it based on Kirchhoff's plate theory.

Fig. 10a, 10b, 10c, 10d, 11a, 11b, 11c .

4.2.2 Large Rotations

An error analysis for nonlinear problems is generally more complicated, sinc the uniqueness of the solution is not guaranteed and branching can occur. To avoid these problems, we restrict ourselves to the stable part of the solution path (below the first critical point) and concentrate on a local error analysis. This can be done by evaluation of the error indicators for linear problems at the linearized equation. For geometrically nonlinear problems with small strains and large rotations, where no a-priori convergence analysis is available, we adopt the simpler approach of controlling the linearization assumptions in the nonlinear formulation.

The relative rotation in the element is given by

$$\mathbf{R}^D = \mathbf{R}(\mathbf{R}^C)^T, \tag{32}$$

where $\mathbf{R}^C$ is defined by the displacements of the element nodes.

In the kinematical equations, the rotation tensor

$$\mathbf{R}^D = \mathbf{I} + \Psi + \frac{1}{2}\Psi\Psi + ... \tag{33}$$

is approximated by

$$\tilde{\mathbf{R}}^D = \mathbf{I} + \Psi \tag{34}$$

in each element.

Thus, the magnitude of the relative rotation $\|\tilde{\mathbf{R}}^D - \mathbf{I}\|$ must be controlled.

From

$$\tilde{\mathbf{R}}^D = \begin{pmatrix} 1 & \psi_3 & -\psi_2 \\ -\psi_3 & 1 & \psi_1 \\ \psi_2 & -\psi_1 & 1 \end{pmatrix} \tag{35}$$

we have (with the norm $\|\mathbf{A}\| = \sqrt{\sum_{i,j=1,3} a_{ij}^2}$)

$$\|\tilde{\mathbf{R}}^D - \mathbf{I}\| = \sqrt{2}\sqrt{\psi_1^2 + \psi_2^2 + \psi_3^2} \overset{!}{\leq} \overline{\eta}_{NL} \tag{36}$$

$$indicator: \ \eta_{NL} = max\ \|\tilde{\mathbf{R}}^D - \mathbf{I}\| \tag{37}$$

This indicator can be directly interpreted as the distortion of the element and is therefor similar to the 'warpage' criterion introduced earlier as an a-priori indicator. Again, a threshold $\overline{\eta}_{NL}$ is defined and all elements with $\eta_{NL} > \overline{\eta}_{NL}$ are refined. Normally $\overline{\eta}_{NL}$ is choosen in the range (0.,0.1). It should be emphasized, that it controls only the linearization assumption; no error measure is associated with it.

Additionally, the jump of the stress resultants can be evaluated as in the linear case. The stresses should be measured in their corotational frame and not transformed to a common coordinate system. The indicator η_{NL} is applied in the nonlinear analysis of an L-shaped beam (Fig. 12, see also [3]).

Fig. 12a,12b

4.2.3 Elasto-plastic problems

Due to the lack of a satisfactory a-priori analysis for the elasto-plastic finite element problem, reliable a-posteriori error estimators are unavailable. Furthermore, linear criteria are not significant because of the stress projection.

We have to resort to a heuristic criterion. For the proper choice of this criterion, the aim of the elasto-plastic analysis must be taken into account. We are especially interested in the ultimate load or failure analysis of structures (possibly including cyclic plasticity).

In this case, the plastic deformations, that cause the failure of the structure, are normally resticted to a small part of the system. It is reasonable to improve the accuracy at the boundary of the plastic zone, which leads to the heuristic indicator

$$\eta_{PL} = J_2 \tag{38}$$

An element is refined, if

$$\eta_{PL} > \kappa\sigma_V \quad , \tag{39}$$

where κ is a constant $\in (0.9, 1)$.

For the analysis of forming process, where large parts of the structure undergo (possibly large) plastic strains, this criterion is not suitable. (This would also require a different mechanical formulation.)

Fig. 13a, 13b

This criterion is demonstrated for an stretched strip with a circular hole (Fig. 13a - c).

5. Algorithms

The finite element discretization of boundary value problem results generally in large systems of linear or nonlinear equations that have to be solved numerically. The proper choice of the solution method strongly influences the obtainable overall accuracy. We first mention that the numerical solution process is - due to round-off - another source of error. These round-off errors are the more serious the larger the condition number of the matrix of the (linearized) system is, which - in turn - grows with increasing number of degrees of freedom and with the approach of critical points on the solution path.

As we restrict ourselves to subcritical loads and stable discretization methods, round-off problems are not significant for moderately large problems. (We have investigated systems with up to $3 \cdot 10^4$ unknowns.) In this range the computational complexity of the solution algorithm is far more important for the obtainable accuracy. In the sequel we therefore discuss efficient methods for the solution of large linear systems of equations

$$Au = f \ , \qquad u, f \in I\!R^n \ , \qquad A \in I\!R^{n \times n} \quad \text{symmetric positive definite,}$$

as they arise from the finite element discretization and give a short outlook on the treatment of nonlinear equations. We are especially interested in the time and space requirements of these methods.

5.1 Direct and iterative methods for linear problems

Elimination methods make use of the positive-definiteness and the band (or skyline) structure of the matrix A. With optimal node numbering, the decomposition time normally increases like $O(n^2)$, and the memory requirements grow superlinearly as well ($O(n^{\frac{3}{2}})$). They are very efficient for small systems, and standard software is available for the assembly and decomposition of the matrix. Furthermore, they can be efficiently vectorized. Nevertheless, the time and space requirement of the linear solver will eventually determine the maximum problem size.

Iterative methods generally have less space requirements (O(n)) due to indirect addressing. Popular methods are Gauß - Seidel - relaxation and it's variants (Jacobi-iteration, successive relaxation). Although there is no general theory available the analysis of a model problem (see [25]) yields the following estimates:

Jacobi-iteration and SOR with $\omega = 1$	: $O(n^2)$ operations
SOR with $\omega = \omega_{opt}$	: $O(n^{\frac{3}{2}})$ operations

In the general case one has to resort to numerical experiments for a sufficient approximation of ω_{opt}. Our numerical tests indicate that even with near optimal ω, SOR is not competitive with decomposition methods for $n < 10^5$.

The method of conjugate gradients (cg) is more promising. The basic convergence estimate for the cg-method is (see [26] e.g.).

$$\| u^k - u \|_1 \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \| u^o - u \|_1 \tag{40}$$

where κ is the condition number of the matrix A and $\| u \|^s := u^T A^s u$. Without further improvements this yields $O(n^{\frac{3}{2}})$ equations for a fixed error reduction.

Acceleration can be achieved by preconditioning: substitute A by $V^{-1} A V^{-T}$ with a suitable matrix V such that $\kappa((VV^T)^{-1} A)$ is smaller than $\kappa(A)$. Popular preconditioning methods are SSOR with optimal ω and uncomplete factorization of A, which reduces - in certain cases - the number of operations to $O(n^{\frac{5}{4}})$ (see [3]). A far more effective preconditioner is the multigrid method, (see below).

5.2 The multigrid method for linear problems

The multigrid method is motivated by an observation about the classical relaxation methods: the long wave components, which correspond to the small eigenvalues of the matrix A, cause the slow convergence of relaxation methods, whereas the highly oscillatory terms converger fast. So, it's useful to perform only a few relaxations steps for the oscillatory terms and approximate the remaining error on an coarser mesh. The proper combination of these two procedures – smoothing step and coarse grid correction – characterizes the multigrid method.

The multigrid method requires a sequence of FE-meshes ${}^{(1)}\mathcal{T} ... {}^{(m)}\mathcal{T}$ and the corresponding spaces of test functions $S_1 \subset ... \subset S_m$. The linear equations on these meshes are ${}^{(l)}A \quad {}^{(l)}u \quad = \quad {}^{(l)}f$. The mesh ${}^{(k+1)}\mathcal{T}$ is constructed from mesh ${}^{(k)}\mathcal{T}$ by subdividing all or - in the case of adaptive refinement - some elements on ${}^{(k)}\mathcal{T}$ into four new elements, (and by introducing special refinement schemes in the transition area in the case of adaptive refinement).

Fig. 14: Graphical representation of 3-grid methods

In the sequel, we give only an outline of convergence results of the multigrid method; the reader is referred to the book of Hackbusch [27] for details about convergence estimates.

An optimal convergence rate can be established by an estimate for the contraction number ρ^k:

$$\rho^k = \frac{\| u^k - u \|}{\| u^{k-1} - u \|} \leq a < 1 \tag{41}$$

where a is - in contrast to the classical relaxation methods - independent of the meshsize. Neglecting post-smoothing, one has for a two-grid iteration:

$${}^{(2)}\mathbf{u}^{k+1} - {}^{(2)}\mathbf{u} = ({}^{(2)}\mathbf{A}^{-1} - \mathbf{P}\,{}^{(1)}\mathbf{A}^{-1}\mathbf{R})({}^{(2)}\mathbf{A}\ {}^{(2)}\mathbf{u}^{k,s} - {}^{(2)}f) \tag{42}$$

where $u^{k,s}$ denotes the result of the pre-smoothing step.

One can now combine the following estimates :

$$\text{(i)} \qquad \| u^{k,s} - u \|_2 \leq s(\nu_1)\frac{1}{h^2} \| u^k - u \|_0 \tag{43}$$

where the smoothing function $s(\nu)$ tends to 0 if ν (= the number of relaxation steps) tends to infinity. This estimate can be easily shown for the Jacobi-iteration. The inequality (43) characterizes the smoothing property of the multigrid method.

$$\text{(ii)} \qquad \| u^{k+1} - u \|_0 \leq c^{n^2} \| u^{k,s} - u \|_2 \tag{44}$$

This is the approximation property of finite element spaces.

Combining (i) and (ii), one finally has

$$\| u^{k+1} - u \|_0 \leq cs(\nu_1) \| u^k - u \|_0 \tag{45}$$

independent of the meshsize.

The above estimate is directly applicable to second order problems without singularities and a conforming finite element method. An analogous result for problems with singularities was given by Yserentant [28]. If more than two grids are involved, the W-cycle iteration can be considered as a perturbed two-grid iteration and a convergence estimate follows immediately. For the V-cycle, the estimates have to be refined. A refinement of the technique is also necessary, if the smoothing property cannot be established directly. (This holds especially for the Gauß - Seidel relaxation, which is very efficient as a smoothing procedure in numerical tests, but still lacks a theoretical foundation in the general case.) The convergence proofs can be extended to fourth order problems. The above discussion shows that the convergence of the multigrid method is guaranteed for conforming finite elements methods (and Jacobi relaxation). The efficiency of the method can be further improved in several ways:

(i) The initial iteration vector is obtained by transfer of the solution on the previous mesh (nested iteration).

(ii) The truncation error needs not be much smaller than the approximation error. Combined with (i), this means that often two or three (even one) iterations are sufficient.

(iii) If the contraction number ρ^k is large (> 0.5 e.g.), the multigrid iteration can still be used as preconditioner for a conjugate gradient iteration, which improves performance significantly.

We conclude this section by presenting some results for the membrane and the plate equation.

5.2.1 The multigrid method for the membrane plate equation

For a conforming FE-method (using bilinear shape functions) the standard multigrid theory is directly applicable. The table 2 contains contractions numbers for a two-grid iteration with different parameters (ν denotes the number of pre- and post-smoothing steps).

Tabel 2: Error reduction factors for a two-grid iteration

From the practical point of view, the computing time is more important. Table 3 compares the CPU time (on CDC Cyber 990, in seconds) for a decomposition method with the time for the multigrid iteration for different numbers of unknowns n.

It should be noted, that the hardware situation is rather unfavourable for the multigrid method, as the matrix decomposition was accelerated by a factor ~ 10 through vectorization, whereas the relaxation and smoothing procedures could not be vectorized. On a typical workstation, the multigrid method would probably have an even bigger advantage over the decomposition method.

In some cases - especially if the coarse grid solution is very inaccurate - the contraction numbers are much larger. This has been identified as a locking problem [30] and remedied by a scaling factor. Our investigations show, that a conjugate gradient iteration - preconditioned with unmodified multigrid cycles - is equally effective. This preconditioned conjugate gradient method can be generally recommended as an efficient and robust algorithm.

Tabel 3: CPU - time for a two-grid method versus decompositions for different problem sizes

5.2.2 The multigrid method for the Kirchhoff plate equation

For this fourth order equation, nonconforming elements are normally used, which makes the analysis of the multigrid method more difficult. Nevertheless, good convergence results have been obtained for some elements (e.g. discrete Kirchhoff theory elements, Morley's element (see [31])).

5.2.3 The multigrid method for the Reissner/Mindlin plate equation

If we use the bilinear shape functions without modification, the multigrid method could be directly applied. The measures against shear locking, however, generate special problems for the multigrid algorithm. Numerical tests with the assumed-strain element [3] exhibit fast convergence for thick plates, a contraction factor of $\approx 0,88$ (with pcg smoothing: ≈ 0.5) for $\frac{h}{d}$ (= elementsize / plate thickness) = $\frac{100}{16}$ and divergence for $\frac{h}{d} > 10$. This problem has not yet been solved. Results for similar elements suggest the use of an equivalent mixed formulation.

5.3 Multigrid methods for nonlinear problems

Multigrid ideas can be used in several ways to speed up nonlinear iterations. In combination with Newton's method, efficient start vectors for the iteration can be obtained by prolongation of a coarse grid solution (nested iteration) and the linear system of equations can be solved by a multigrid iteration (Newton-multigrid method, Newton MG). It is also possible to apply the multigrid method with nonlinear smoothing directly to the nonlinear problem (full approximation scheme, FAS). A comparison of the Newton MG with FAS applied to a geometrically nonlinear membrane and plate bending problem shows that the convergence of the FAS often is better than that of the Newton MG, if the nonlinearity gets a stronger influence on the solution. The full approximation scheme needs less memory, but more time for the smoothing step than the Newton MG.

As in the linear case both methods can be accelerated by a line search or a combination with a conjugate gradient scheme (see [31]). A combination of the two methods, starting with nested iteration, applying some steps of the FAS and then, in the near of the solution, changing to the Newton MG with several linear multigrid cycles for one linear system of equations is especially promising.

6. Mesh generation and refinement

We first give an survey of the whole process and then discuss single topics like the definition of a structural model, the generation of an initial mesh and the refinement of the initial mesh in more detail.

6.1 Survey

In the practical application of FE-methods the generation of a suitable FE-mesh for the structure under consideration is often extremely time consuming. At present a number of commercial mesh generators are available which simplify the generation process; but one is far away from an entirely automatic procedure which is working with a standard interface to CAD-data.

For a combined application of mesh refinement and multigrid methods none of the standard programs meet the requirements. Thus a new concept has to be developed. This concept is adjusted to an automated discretization and calculation within specified tolerances and without further actions by the user. It distinguishes between four stages in the mesh-generation process which are mainly independent from each other.

At first one presupposes that a representation of the structural model is given. This structural model either contains references to a corresponding CAD-representation or an explicit description of the geometry. This description of the geometry has to be enriched by the quantities that are needed to formulate the boundary value problem (material properties, loads, boundary conditions, kinematics...).

Based on this structural model an initial mesh in constructed, i.e. the structural model is split in three- and four-sided regions and the topological relationship between the regions is determined.

In a third step a mesh, that is suitable for a FE analysis, is generated by refinements on the basis of a-priori criteria. This "admissible mesh" can be refined on the basis of a-posteriori criteria and transforms then to a mesh, that is quasi-optional with respect to the ratio of cost versus accuracy. This mesh is called in following "adapted mesh". On that mesh one should be able to perform a multigrid calculation. These steps can be seen in Fig. 15 a) -d)

Fig. 15 a) - d) : Steps of mesh generation and refinement

6.2 Structural model

A description of the structural model by means of a normed data interface is desirable. Unfortunately the customary data-interfaces as VDA-FS, IGES, SET are not in the position to do this at present. Nevertheless there are activities to supply this want; we refer to CAD*I and the ISO-norm STEP which are under development.

In this situation a new data model was developed to implement the algorithms in INA-SP (INA-SP : Inelastic Analysis of Shells and plates, developed at IBNM, University of Hannover) and ICEM SURF (ICEM SURF : developed by Control DAta Corp.). A relational concept was chosen, using objects and relations between these objects. Any kind of attribute can be assigned to objects and their relations.

The database structure does not only make the common distinction between the dimensions of the object, but has in addition the capability of a hierarchical differentiation to support surface or part-oriented operations. In addition, topological and geometrical data can be distinguished, as well as data used in different models. It is an open concept allowing a connection with other models and hierarchies.

6.3 Initial mesh

In the following we will distinguish between initial mesh generation and " mesh refinement". The task of initial mesh generation is to deliver a first discretization of a structural model. That means, an initial mesh generator disintegrates a structural model, which is essentially given by purely geometrical quantities (possibly supplemented by nominal topological specifications), into a number of three- and four-sided "elements " (in the case of free-form surfaces). These elements are connected with their neighbours by the numbering of their nodes and edges. The topological connection by the numbering of edges is needed since there is no guarantee that the initial mesh is conforming. A mesh is conforming, if the interaction of any two adjacent elements is either a vertex or an edge of both elements. Nodes violating this condition are called irregular nodes. The automatic generation of the topology out of purely geometrical information should be possible for branching and overlapping structures. The generated elements have to be independent of the representation of the structural model.

The simplest possibility to produce an initial mesh is the enrichment of the structural model by all the topological quantities needed, i.e. by the information about nodes, edges and by manually collecting nodes and edges to elements. This way was chosen in INA-SP.

A second possibility is to divide the structural model into simple components (e.g. simply connected regions, regions with three or four distinguished boundaries). The topology of these boundaries has to bee specified explicitely. The partition into triangles and quadrilaterals could then be done in an automatic way. Common procedures for this task rest on, e.g. the mapped meshing principle [32], the Dirichlet tesselation [33], the best splitting method [34] and the quadfree method [35].

The next step should be the generation of the topological relation between the simple components automatically. For regions with three or four distinguished boundaries this was done in ICEM MESH [36]. It should be pointed out that if T-connections are present, a node topology is not sufficient to correctly identify and process branched structures and topological relations between adjacent elements. This requires the more efficient segment topology.

Fig. 16 : Transitition from CAD-geometry to CAD-topology

6.4 Mesh refinement

The initial mesh is neither necessarily conforming nor is the presence of distorted elements forbidden which are rejected by many inite element programs, nor must these elements represent a meaningfull approximation of the structural model. The refinement algorithm has to guarantee the mentioned properties as well as the quasi-optimality of the mesh. It is to be ensured that local criteria are met by local refinements, without any propagation further than to the adjacent elements, without any reduction of angles, and generally producing quadrilateral elements.

The technique used here is a recursive application of the mapped meshing concept. As criteria for the refinement serve the nonconformity of element edges, the a-priori and the a-posteriori criteria described in chapter 4. While a-posteriori criteria cause a homogenous refinement of an element, the refinement scheme caused by a-priori criteria or by the nonconformity, respectively depends on the criterion, in order to generate new elements which fit better to the criteria than the original one.

In common algorithms [22] the adaptation of neighbouring elements to a refined element causes difficulties. One could circumvent these difficulties by generating new nodes in the interior of the original element only. But the new elements show unfavourable angles compared with the original elements. After some refinement steps the mesh would entirely useless. Thus it is an essential goal to preserve the shape of elements or to improve it respectively. This can only be done by subdividing the elementedges. Usually the elementedges are halved. It is not recommended to generate more than two edges out one element edge, since the number of nodes is rapidly increasing, the multigrid methods do not work efficiently and the algorithm will be substantially more complex.

By the refinement of elementedges the problem of conforming elements arises. If irregular nodes are retained - as it was suggested by [37], [22] - the degrees of freedom at these nodes need to be eliminated in advance. The drawback of this approach is, that one has to interrere with commercial FE-programs usually is not possible. For conforming meshes the algorithms are considerably larger complixity. The method used here is called "shape preserving recursive mapped meshing" (SPRMM) and will be described in the following.

After the evaluation of all adaption criteria, the following situation occurs at the beginning of a refinement step: for a portion of the elements some or all of the edges are marked for refinement. Frequently inadmissible refinement patterns arise (e.g. by propagation of refinements in adjacent elements), usually there are quadrilaterals with three edges to be refined. In this case the fourth edge will be refined, too. This adaption step for the edges will be repeated until only admissible refinementpatterns appear. In Figure 17 some of these patterns and the corresponding refined elements are displayed.

(a) (b) (c) (d)

Fig. 17 : Some refinement patterns and the corresponding elements

6.5 Refinement of modified elements

Problems are caused by the refinements (c) and (d) in Fig. 17. These refinements produce elements which halven an angle in the original element. Such refinement will be called "modified". Modified elements violate the goal not to deteriorate the shape of the original elements. On the other side, there is on way to generate conforming and adaptive meshes without halvening angles.

In SPRMM it is presupposed that the angles are chosen in such a way that one (and only one) division of an angle is allowed. Therefore it is necessary to prevent the angles from a repeated subdivision. Hence modified elements are always treated in a special way. A-priori criteria with respect to the element shape are not applied to them. If an additional subdivision is required by the adaption to the refinement in an adjacent element or by other criteria, one goes back to the unmodified original element and eliminates the modified edge. This procedure is demonstrated by the following example (see Fig. 18). The pattern (d) in Fig 17 has to be refined since two adjacent elements have been subdivided. After elimination of the modified edges in mesh 2 all the elements in mesh 3 are regular.

mesh 1 mesh 2 mesh 3

Fig. 18 : Refinement of a modified element

Thus it is ensured that angles are deteriorated only by a factor of 0.5. The programming of this algorithm is quite expensive, since a lot of special cases have to be taken into account.

6.6 Remarks

- With the respect to the combination with multigrid methods it should be noted that the inclusion of the spaces of shape functions is violated locally by the elimination of modified edges.
- For multigrid methods the nearly generated nodes are entered into the element- and edge-tables of the new level. Since the kind of shape functions for the elements and edges are stored in a table too, there are no additional information required to proceed with the transfer.
- The refinement algorithm is independent to a large extend from the type of the shape functions and therefore from the number of nodes in one element. This number is only

important for the node generation; at present in addition to the three- and four-noded elements with six and nine nodes are realized. A further increase of the number of shape functions or a different number of nodes on element edges makes no difficulty.

- Some examples of the application of the refinement algorithm to complex real life structures are shown in the following figures 19 - 21, where different steps of this process are shown. In part a) of those the initial FE-mesh is plotted and in part b) the refined mesh reguarding all a-priori criteria. We thank the Volkswagen AG, Wolfsburg; Aerospatial, Toulouse and Control Data Coporation, Minneapolis, for permitting the publication.

Fig. 19 : Roof girder of a motor car
19 a) : Initial FE-mesh
19 b) : Shaded initial FE-mesh
19 c) : Detail of the refined mesh

Fig. 20 : Outer body of a motor car (Audi Coupe)
20 a) : Structural model
20 b) : Initial FE-mesh
20 c) : Refined FE-mesh

Fig. 21 : Suspension of the engine of an aircraft (Airbus)
21 a) : Structural model
21 b) : Initial FE-mesh
21 c) : Detail of the suspension of the engine

7.Conclusion

A concept for the coupling for CAD and FEM by automatic mesh generation for widely arbitrary criteria was presented where the a-priori criteria were: element size, element shape, deviation from the given geometry and the conformity of the elements. The a-posteriori indicators for mesh refinements are besides the well-known estimators for linear elastic problems, a control of the linearization assumptions for geometrically nonlinear problems and a heuristic refinement criteria derived from the von Mises J_2- yield condition.

The method for the construction of conforming finite element meshes is independent of a specific CAD data representation. The tools are an automatic topology generation, an appropriate database, refinement criteria and a local refinement strategy. In total, incorporating the computational experience, the presented integrated adaptive engineering system have proven to be reliable for the mechanical modelling and the discretization and efficient with respect to the solvers.

References

[1] P. G. Ciarlet and P. Destuynder. A justification of the two-dimensional linear plate model. *Journal de Mécanique*, 18:315–344, 1979.

[2] M Bernadou, S. Fayolle, and F. Léné. Numerical analysis of junctions between plates. *Computer Methods in Applied Mechanics and Engineering*, 74:307–326, 1989.

[3] E. Stein, K.-H. Lambertz, and L. Plank. Error estimators and mesh adaptation for thin-walled structures at finite rotations. In G. Kuhn and H. Mang, editors, *Proceedings of the IUTAM Symposium "Discretization Methods in Structural Mechanics", Wien 1989*, to appear in Springer-Verlag, 1990.

[4] K.-H. Lambertz. *Traglastberechnungen von Faltwerken mit elastoplastischen Deformationen.* Technical Report F 89/1, Forschungs- und Seminarberichte aus dem Bereich der Mechanik der Universität Hannover, 1989.

[5] K. J. Bathe and E. Dvorkin. A continuum mechanics based four-node shell element for general nonlinear analysis. *Engineering Computations*, 1:77–88, 1985.

[6] K. J. Bathe and F. Brezzi. On the convergence of a four-node plate bending element based on Mindlin/Reissner plate theory and a mixed interpolation. In *The Mathematics of Finite Elements and Applications*, 1985.

[7] D. N. Arnold and R. S. Falk. *A Uniformly Accurate Finite Element Method for the Mindlin-Reissner Plate.* Technical Report 307, IMA Preprint, 1987.

[8] D. N. Arnold, F. Brezzi, and J. Douglas Jr. PEERS — A new mixed finite element for plane elasticity. *Japan Journal of Applied Mathematics*, 1:347–367, 1984.

[9] E.Stein and R. Rolfes. *Mechanical conditions for stability and otimal convergence of mixed finite elements for linear plate elasticity.* Technical Report 89/6, Universität Hannover, Institut für Baumechanik und Numerische Mechanik, 1989. Accepted for publication in *Computer Methods in Applied Mechanics and Engineering.*

[10] I. Babuška, R. B. Kellogg, and J. Pitkäranta. Direct and inverse error estimators for finite elements with mesh refinements. *Numerische Mathematik*, 33:447–471, 1979.

[11] G. Strang, H. Matthies, and R. Temam. Mathematical and computational methods in plasticity. In S. Nemat-Nasser, editor, *Variational Methods in the Mechanics of Solids, 1978*, Pergamon Press, 1980.

[12] C. Johnson. On finite element methods for plasticity problems. *Numerische Mathematik*, 26:79–84, 1976.

[13] M. Ortiz and E. P. Popov. *Accuracy and Stability of Integration Algorithms for Elastoplastic Constitutive Equations.* Technical Report, Division of Engineering, Brown University, Providence, 1984.

[14] E. Stein, K.-H. Lambertz, and L. Plank. Verfahren zur Traglastberechnung dünnwandiger Strukturen mit elastoplastischen Deformationen. In E. Stein, editor, *Nichtlineare Berechnungen im Konstruktiven Ingenieurbau*, Springer-Verlag, 1989.

[15] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland Publishing Company, 1976.

[16] J. Robinson and G. W. Haggermacher. Element warning diagnostics. *Finite Element News*, 3:30–33, 1982.

[17] J. Robinson and G. W. Haggermacher. Element warning diagnostics. *Finite Element News*, 4:19–23, 1982.

[18] I. Babuška and E. Rank. An expert-system-like feedback approach in the hp-version of the finite element method. In *Finite Elements in Analysis and Design 3*, 1987.

[19] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM Journal on Numerical Analysis*, 15:736–755, 1978.

[20] I. Babuška and W. C. Rheinboldt. Computational error estimates and adaptive processes for some nonlinear structural problems. *Computer Methods in Applied Mechanics and Engineering*, 34:895–937, 1982.

[21] I. Babuška, D. W. Kelly, J. P. de S. Gago, and O. C. Zienciewicz. A posteriori error analysis and adaptive processes in the finite element method - part i: Error analysis. *International Journal for Numerical Methods in Engineering*, 19:1593–1619, 1983.

[22] E. Rank. *A-posteriori Fehlerabschätzung und adaptive Netzverfeinerung für Finite-Element- und Randintegralelement-Methoden.* Technical Report, Institut für Bauingenieurwesen, Technische Universität München, 1986.

[23] E. Stein,K.-H. Lambertz, L. Plank, and A. Reisch. *Elementbenchmark Rhombusplatte 60° : Ergebnissse für DKT-Dreieckselement, DKT-Viereckselement, Mindlin-Element mit reduziertem Schubansatz.* Technical Report 87/12, Universität Hannover, Institut für Baumechanik und Numerische Mechanik, 1987.

[24] P. S. Theocaris and E. Marketos. Elastic-plastic abalysis of perforated thin strips of a strain hardening material. *J. Mech. Phys. Solids*, 12:377-391, 1974.

[25] D. M. Young. *Iterative Solution of Large Linear Systems.* Academic Press, 1971.

[26] O. Axelsson and V. A. Barker. *Finite Element Solution of Boundary Value Problems: Theory and Computation.* Academic Press, 1984.

[27] W. Hackbusch. *Multi-Grid Methods and Applications.* Springer-Verlag, 1985.

[28] H. Yserentant. The convergence of multi-level methods for solving finite-element equations in the presence of singularities. *Mathematics of Computation*, 47:399–409, 1986.

[29] D. Braess. A multigrid method for the membrane problem. *Computational Mechanics*, 3:321–329, 1988.

[30] D. Braess and P. Peisker. A conjugate gradient method and a multigrid algorithm for Morley's finite element approximation of the biharmonic equation. *Numerische Mathematik*, 50:567–586, 1987.

[31] W. Rust. Mehrgitter-Verfahren für FE-Formulierungen geometrisch nichtlinearer Scheiben- und Plattenprobleme mit Konvergenzbeschleuniger. *to appear in ZAMM*, 1990.

[32] O. C. Zienciewicz and D. V. Phillips. An automatic mesh generation scheme for plane and curved surfaces by 'isoparametric' coordinates. *International Journal for Numerical Methods in Engineering*, 3:519–528, 1971.

[33] J. C. Cavendish. Automatic triangulation of arbitrary planar domains for the finite element method. *International Journal for Numerical Methods in Engineering*, 8:679–696, 1974.

[34] M. L. C. Sluiter and D. L. Hansen. A general purpose automatic mesh generator for shell and solid finite elements. In L. E. Hulbert, editor, *Computers in Engineering 3*, pages 29–34, ASME, 1983.

[35] M. S. Shephard, M. A. Yerry, and P. L. Baehmann. Automatic mesh generation allowing for efficient a priori and a posteriori mesh refinement. *Computer Methods in Applied Mechanics and Engineering*, 55:161–180, 1986.

[36] D. Bischoff. ICEM VWMESH — A mesh generation tool for free-form surfaces. In *Proc. Conf. StruCoMe, Paris, 1989*, DATAID, 1989.

[37] G. F. Carey. A mesh refinement scheme for finite element computations. *Computer Methods in Applied Mechanics and Engineering*, 7:93–105, 1976.

Framework for the Reliable Generation and Control of Analysis Idealizations

Mark S. Shephard, Peggy L. Baehmann
Marcel K. Georges and Elaine V. Korngold

Program for Automated Modeling
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

1. SUMMARY

This paper examines the sources of idealizations commonly employed during the application of engineering analysis, and indicates the techniques available to control the errors they introduce. A framework is presented of a modeling system that can support the various levels of analysis idealization control. Methodologies for the use of such a system in the design and analysis of airframes and plane elasticity problems are discussed. Finally, comparisons of the computational efficiency of automated, adaptive analysis techniques to control the discretization error in the energy norm for plane elasticity problems are presented.

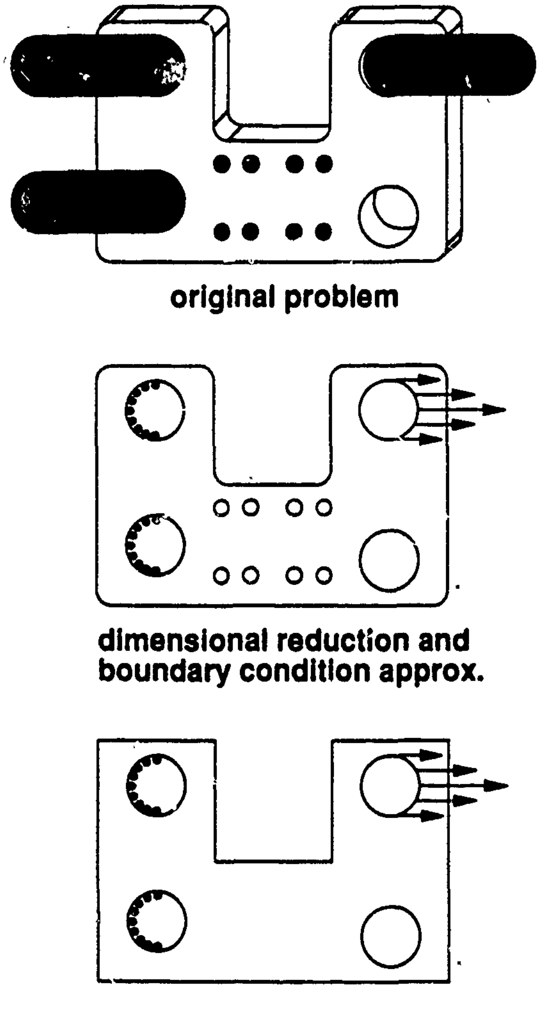
2. INTRODUCTION

The area of computational mechanics has advanced to the point that it is possible to perform reliable engineering analyses for a wide variety of physical problem types. The availability of these capabilities in the form of general purpose analysis codes that run on fast and inexpensive computers has provided engineers with the ability to perform various forms and levels of analyses during the design process. However, the ability of the design engineer to reliably apply the current general purpose analysis tools to determine a given set of performance parameters is questionable. In general it is not possible to provide every design engineer with the extensive training needed in the techniques underlying the analysis tools to reliably apply those tools. It is therefore necessary to continue to develop the tools to provide engineers with the needed capabilities. These tools must explicitly control all levels of idealization associated with each analysis performed.

There is a tendency of some to assume that the development of this set of tools is not an appropriate goal, because every engineer should fully understand every tool they use. In today's environment this is no longer a practical point of view. The number and sophistication of the tools available and required to do engineering design in a competitive manner is beyond the ability of the engineer to learn in any acceptable time frame. It is more important for engineers to have the training needed for the tasks they are going to perform, and that the tools provided to them do reliably perform the tasks they are claimed to be capable of performing.

Idealization errors in engineering analysis arise from a number of sources ranging from the selection of the mathematical model for the physical behavior of interest, to the discretization errors associated with a finite element mesh.The techniques available to control the errors introduced by these idealizations range from simple rules based on experience to bounded a posteriori error estimates. It is important that techniques that improve the ability to reliably control the errors due to idealization continue to be developed. It is also important that the tools necessary to support all levels of idealization control be integrated with design modeling systems.

This paper discusses the idealizations common to engineering analysis and the techniques available to control those idealizations. It then examines the framework of a modeling system that can support the various levels of idealization control and how two specific groups of idealizations may be controlled using such a system. Finally, the computational efficiency of some current procedures for the adaptive control of the idealization errors due to finite element mesh discretization are given.



original problem

dimensional reduction and boundary condition approx.

geometric simplification

Figure 1. Idealization of simple bracket

3. IDEALIZATION IN ANALYSIS AND DESIGN

The application of engineering analysis typically employs a number of idealizations to reduce a physical behavior to a set of algebraic equations that can be solved manually or on a computer. Figure 1 depicts some of the idealization processes typical for reducing a simple mechanical problem to a form appropriate for numerical analysis. Each step of idealization used in an engineering analysis process introduces some level of approximation error. The reliability of an analysis depends on the ability to understand and control the approximation errors introduced by each step of idealization [1], [2], [3], [4].

The accuracy of a solution is a function of the measure(s) of accuracy you are interested in. When an analysis is performed, there are several parameters of interest for which error control is desired. Therefore, the goal of the analysis is to obtain a solution such that

$$e_i \leq \Upsilon_i \qquad\qquad i = 1,2,3,...,m$$

where e_i is the value of the error for the ith norm, Υ_i is the limit allowed for the error in the ith norm, and m is the number of error norms being considered.

The first step to estimating the errors in engineering analysis is to enumerate the contributing sources as

$$e_i = e_i(\Psi, \Omega, \psi, D, \beta, \Delta)$$

where Ψ is the base mathematical model selected to represent the physical behavior of interest, Ω is the domain of the analysis, ψ are the dimensional reductions and associated alterations of the mathematical model to eliminate physical dimensions, D are the material property parameters, β are the boundary conditions (also initial conditions when time is one of the dimensions of the problem), and Δ represents the discretization used for the analysis.

The brief descriptions given below indicate what each source of error is. A more detailed discussion of these error sources including the example of their relation to the analysis of a reinforced concrete dome roof are given in reference [3].

Mathematical model, Ψ. The derivation of a mathematical model employs physical laws, mathematical manipulations, and behavioral assumptions. The behavioral assumptions are needed so that the physical laws can be mathematically manipulated to yield a useful set of expressions. Each behavioral assumption introduces approximation, and associated error. Therefore, the errors introduced by each assumption must be considered.

Domain, Ω. There is no additional error introduced into the analysis if the domain is complete with respect to the mathematical model being solved. The domain used in most analyses is typically not complete in that it may be limited to a portion of the total domain, with boundary conditions applied along the boundary introduced by the eliminated portion of the domain. If the portion of the domain eliminated is symmetric to that analyzed, and the proper essential boundary conditions are applied, there is no error introduced into the analysis. In many cases the portion retained is not symmetric with the portion eliminated. In these cases, an error due to geometric approximation is introduced. It is also common to eliminate geometric details which also introduces errors into the solution.

Dimensional reductions and associated mathematical model alterations, ψ. It is often possible to selectively introduce specific behavioral assumptions over portions of the domain which allow the simplification of the mathematical model by reducing the physical dimensionality of that portion of the domain. One example that has received considerable attention, and is the subject of several of the papers in this volume, is the elimination of one dimension from the three-dimensional equations of elasticity to produce plate and shell formulations. The eliminated physical dimensions are accounted for by the introduction of specific parameters. The application of dimensional reductions is an additional source of solution error.

Material properties, D. The mathematical model, Ψ, as well as dimensional reduction and associated mathematical model alterations, ψ, fix the framework of the material model. However, within that framework there is still the specification of the parameters. Any variance from the correct form and values introduces error into the solution.

Boundary and initial conditions, β. The mathematical model, Ψ, as well as dimensional reduction and associated mathematical model alterations, ψ, specify the boundary and initial conditions that are part of the current analysis. The boundary and initial conditions which must be specified are often difficult to abstract from the physical situation being analyzed. In addition, the values of many of them, particularly natural boundary conditions, are probabilistic in nature. Therefore, they can be another source of error in the analysis.

Discretization, Δ. Discretization consists of representing the continuous field by one written in terms of a specific number of parameters. This reduction of the solution test space introduces approximations into the analysis process.

Since the exact solution to a requested analysis is generally not known, it is only possible to obtain estimates to the solution error. The goal of idealization error control is to ensure that

$$E_i \leq \Upsilon_i \qquad\qquad i = 1, 2, 3, \ldots, m$$

where E_i is the value of the error estimate for the $\mathrm{i^{th}}$ norm, $E_i \rightarrow e_i$ as the solution procedure is refined, and $E_i \simeq e_i$ for a solution of acceptable cost.

The techniques available to aid in the control of idealization errors include:

1. Analytically-based error estimation
2. Analytically-based results for ideal situations
3. Model improvement through hierarchic model comparisons
4. Sensitivity analysis
5. Statistical methods
6. Comparison to known physical limits
7. Comparison to test results
8. Comparison to reasonable limits
9. Rules based on experience and intuition

The ability of these techniques to reliably control idealization errors varies greatly. However, it is not always possible to derive an entirely reliable control method for many idealizations. Therefore, the preferred method of controlling idealizations is to use the best method available for each, and to strive to develop improved methods to control those idealizations for which control can be improved.

Analytically-based idealization control techniques based on bounded error estimates are the most reliable because they provide a direct measure of the error contribution. The development of a posteriori error estimators to control the idealization errors due to discretization is currently an area of active research and development [5], [6].

Another method to improve the reliability of a solution process is to employ analytically-based results for idealized situations within a more general analysis process. The development of these types of procedures is common place for the solution of problems where behavioral considerations at two different size scales are critical to the prediction of the behavior at both scales. A common class of problems are materials constructed from multiple constituents including composite materials [7], [8], [9], soils [10], and hydrated tissues [11], [12]. In this class of problem the micromechanical behavior of the individual constituents determines the overall behavior of the material. A number of homogenization techniques [2], [7], [8], [11], [10], [9] have been developed to define a set of homogeneous material properties for a macromechanical level analysis that accounts for the micromechanical behavior of the constituents. Such techniques may also be useful in the development of procedures for supporting specific classes of geometric simplifications.

Model improvement through hierarchic modeling comparisons involves solving a model at one level of representation and then using that information with a hierarchically better representation to determine any required improvements in the current solution as well as providing additional local information. Noor [13], [14] has developed such a procedure for the analysis of composite plates and shells. This procedure performs a macromechanical analysis using a lower order transverse shear representation. The overall solution parameters at specific locations are used to define the information for an approximate local three-dimensional analysis. The three-dimensional calculations give useful estimates of local parameters as well as an improved shear correction parameter for use to obtain improved values of the overall solution parameters.

The application of sensitivity analysis methods does not give a direct measure of the error due to an idealization. Their primary use is to give a measure of how much variation in a specific set of performance parameters can be expected from variations of another parameter. For example, the sensitivity analysis procedures used in shape optimization [15] can provide useful information on the sensitivity of many structural analysis solution parameters to changes in the model domain definition. Sensitivity analysis information can be used to improve the reliability of engineering analysis procedures by indicating the range of variation to be expected in one set of parameters based on other parameters. They can also be used to point out which solution parameters must be most explicitly controlled since their values are subject to large variations.

Statistical methods are important for providing an analysis of the influence of variations of problem definition parameters on the solution parameters. For example, Babuska [1] has analyzed the influence of stochastic loads and boundary conditions on various output parameters such as yield surfaces and fracture criteria. The results demonstrate the need to carefully perform such analyses since a given level of variance on input parameters and their distribution can lead to substantially different variances in the solution parameters of interest.

Comparison to known physical limits and comparison to test results represent relatively simple procedures to check the validity and/or accuracy of an analysis. In both cases the analysis results are compared to information that has been specifically measured. The differentiation made between the two techniques is that physical limits are commonly know and tabulated measures, such as the melting point of a material, while test results are assumed to give specific measures of parameters on a full scale or component test that matches an analysis perfomed. Comparisons to known physical limits typically provide simple validity checks on some basic behavioral assumptions used in the analysis.

Comparisons to physical tests are more commonly used to determine the ability of the analysis procedures to accurately predict the solution to an analysis. The application of such comparisons range from proof testing of the final design to the development of specific modeling technologies. One example area currently under study by Szabo [4] is the modeling of structural connections used in the aerospace industry. Due to the number of such connections and their behavioral complexity, a number of simplified modeling procedures, which have been "verified" and tuned using test result information, have been developed. An interesting result of this investigation is that although there appears to be a correlation with test results, the modeling procedures often applied are incorrect because they employ mathematically inadmissible representations of the support conditions [4].

Comparisons to reasonable limits can be used to help improve the reliability of a solution process in a manner similar to comparing solution results to known limits. The differentiation here is that the reasonable limits are not hard physical limits, instead they represent limits that have been found to provide a reasonable measure of the limits of validity of analysis assumptions. One example is to assume that it is satisfactory to employ a small strain formulation when the strains are below a given limit. Such simple comparisons provide an inexpensive means to check assumptions where the only alternative to determining the influence of the assumption is the expensive process of performing the analysis without that assumption. The values of the reasonable limits are often obtained from previous results which have provided measures of the importance of specific assumptions on the analyses. To be used in a reliable manner, such limits must be conservatively set, with previous demonstration of the generality of employing such limits. The generality of such limits is best determined through appropriate mathematical analysis.

The final technique to improve the reliability of a solution process it to employ rules based on the experience and intuition of those that have successfully defined and applied those rules. Over the years, most industries develop and document specific sets of analysis modeling rules that are appropriate to their specific class of problems. The company's engineers employ, and when appropriate, improve these rule sets for future use. One set of such techniques for the dimensional reductions and discretizations used to develop airframe finite element models is given in [16].

The paragraphs below briefly discuss the idealization control techniques with respect to the idealization error sources.

Mathematical model, Ψ. The primary purpose of idealization error control for a mathematical model is to qualify the behavioral assumptions used to derive that model. The application of analytically-based procedures is difficult since these require the existence of mathematical models to measure against. The exceptions to this would be where a related set of hierarchical mathematical models would be available to measure against, or at least compare to each other. However, it must be realized that this approach requires the existence of a high level mathematical model that is known (assumed) to be correct.

One idealization control method that can be applied at this level is the comparison of the results to known physical limits. Two simple examples of this type are, checking the stress field from a linear elastic analysis against a yield criterion, and comparison of the temperature field from a thermal analysis excluding phase change with the melting temperatures of the constituents. In some classes of analyses such straight forward checks are not as simple. For example, it is often not possible to examine the results of a fluid flow analysis to see if the exclusion of a specific consideration would strongly affect the solution. In such cases, experience (either from experimental comparisons or trial and error) may be useful in providing guidance as to reasonable limits for the application of specific mathematical models. Finally, specifically designed experiments to measure selected critical parameters must often be devised to help test the appropriateness of a mathematical model.

Different levels of mathematical models are often used in the design process to evaluate an aspect of a design. Often early models are used to efficiently give overall estimates of specific quantities. Later analyses use more accurate mathematical models that require more computational effort and detailed design information. Still, other mathematical models include physical behaviors not represented in earlier models. Control and coordination of such series of analyses must focus on the result desired from each analysis, the accuracy of result desired, the state of the design, and the analysis information available from previous analyses.

Domain, Ω. Methods to control idealization due to domain simplifications depend upon how the geometric simplification effects the solution behavior.

The first category of domain simplification that must be identified are those that alter the form of the solution behavior. For example, approximating a smooth boundary with a faceted one can change the exact values of many quantities to a given mathematical model from smooth to singular. Depending on the goal of the analysis such approximations may yield the solution useless [1], or simply indicate that the values to given solution parameters are not meaningful in specific areas [3], [17]. The determination of when these circumstances will arise requires an analytic understanding of the basic solution behavior [1], [2], [18].

A number of techniques are available to estimate and control geometric approximation errors when the approximations do not alter the smoothness of the solution with respect to the parameters of interest. A priori error estimates have been derived for specific classes of geometric approximation [19]. Such estimates can be used as the basis of the development of a posteriori error estimates for these approximations. Sensitivity analysis can be used to determine how important geometric variations may be on solution results. The various techniques for sensitivity analysis from shape optimization [15] provide a set of tools for this situation. Analytically-based results to idealized situations can also provide a useful set of techniques. A specific example of such a technique for estimating the influence of circular holes in stress analysis is given in [20]

Sensitivity analysis methods can also be useful to estimate the influence of replacing a portion of a domain by a given set of boundary conditions that do not strictly adhere to a given symmetry situation. Sensitivity measures to small changes in the boundary conditions approximating the behavior of the eliminated portion of the domain can provide useful information in the control of errors due to this type of idealization. At the simpliest level, a high sensitivity indicates a need to include the eliminated portion of the domain in the analysis. A more appropriate use of such procedures would be to bracket the variations of solution results based on the expected variations of the boundary conditions used to approximate the eliminated portion of the domain.

Dimensional reductions and associated mathematical model alterations, ψ. The techniques available to control dimensional reductions vary greatly in sophistication and cost of application. Since dimensional reduction can typically be stated as the application of additional behavioral assumptions on a given mathematical model, the ability to control the approximation errors depends on the ability to qualify the applied assumptions.

Given a starting mathematical model, the approximation errors introduced by its dimensional reduction can be controlled by analytically-based procedures if the dimensional reduction can be represented as a convergent sequence. One example that has been considered is the dimensional reduction associated with the analysis of plate like domains [21]. The key to this procedure is to not state a specific assumption on the 'through the thickness' behavior, but to discretize it with a convergence expansion.

Since dimensional reduction begins with a mathematical model in a dimension higher that the reduced dimension problem to be solved, there is always the possibility to do comparisons with various levels of models. A simple example is the comparison of various levels of plate theory solutions. A more appropriate application of such a concept is that used by Noor [13], [14] that combined overall results at the global level with local level calculations on a higher order model.

Several of the other idealization control techniques are also useful with dimensional reductions. The various techniques can be used alone or in combination with others.

An example where test results are combined with dimensional reduction is the analysis of structural connections [4]. In this analysis the material the connection bares against is replaced by a set of nonlinear springs where the stiffness parameters of the springs are determined by matching the connection analysis behavior to that of a physical connection that was tested.

Material properties, D. The specification and control of material properties must ultimately rely on the input from physical tests of components designed to measure the parameters required in the selected form of constitutive relationship.

In many cases the form of mathematical model used is dictated by the form of material constitutive model required to capture the needed physical behavior. In many cases analytic analyses indicate the form of constitutive relationship that is meaningful. For example, the theory of plasticity has been strongly influenced by the desire to employ constitutive forms that demonstrate specific mathematical properties.

Sensitivity analysis and statistical analysis have a strong influence on determining the variations of solution parameters based on expected variations in material parameters.

Boundary and initial conditions, β. As with the domain approximations, the methods to control idealization due to boundary and initial condition approximations depend upon how they effect the solution.

The first category of simplification that must be identified are those that alter the form of the solution behavior. For example, approximating a distributed load or boundary condition with a sharp variation can introduce singularities into the solution. These singularities can range from analytic singularities, to non-analytic singularities for which the solution is meaningless [1], [4]. These situations would be handled with the same types of techniques as domain simplification.

The same techniques available for domain simplification are available when the approximations do not alter the smoothness of the solution with respect to the parameters of interest.

Discretization, Δ. The mathematical basis of the discretization process makes it possible to control the discretization errors through the application of a posteriori error estimation techniques [5], [6]. Over the past several years these techniques have been developed and combined with adaptive techniques that automatically control discretization errors through the successive enrichment of the discretization. These techniques have matured to the point that they are beginning to be considered for inclusion in commercial analysis codes [22].

In the long term, it is expected that most discretization errors will be controlled by adaptive analysis techniques. At this time good adaptive procedures to control all discretization errors do not exist, and even when they do, the overall efficiency of the process may be improved by combining them with other techniques. For example, rules based on an analytic understanding of the solution behavior, and/or previous experience can provide a priori information for the development of an initial discretization. Such a capability is needed when there are no adaptive procedures available to control the discretization errors. They can also be useful in conjunction with adaptive techniques to improve the computational efficiency of the analysis process [23], [24].

4. THE GENERATION OF IDEALIZED MODELS

Given a definition of the object of interest and a knowledge of the analysis idealizations to be invoked, the task of generating the idealized model in the form needed for the analysis procedure must be performed. Historically, the view of this process taken by analysts is that the only definition of the object that exists is the idealized model. Considering the broader context of a product design process, this is not appropriate. The description of the object evolves with the design and is never necessarily equivalent at any particular idealized analysis model. For sake of discussion, it is assumed that a computerized representation of the object to be analyzed exists in a design modeling system. Therefore, the generation of the idealized analysis model consists of the following steps:

1. Alteration of the domain and analysis attribute information based on the geometric simplification and dimensional reduction idealizations.
2. Specification of the appropriate constitutive laws based on the material property information.
3. Performance of boundary condition idealizations.
4. Generation of a discretization of the idealized model as needed for the analysis procedure.

The alteration of the object domain as required to reflect domain simplifications and dimensional reductions can be performed by the application of an appropriate set of geometric modeling operations. This can only be done within the context of the modeling system if the geometric modeling tools and representations available can support the required operations. In general this requires a general non-manifold geometric representation that can uniquely house any combination of one-, two, and three-dimensional entities in three space [25], and a complete set of non-manifold geometric operators [26].

To be performed in a reliable manner, the geometric modeling operations required to create the idealized analysis model domain must operate in an automatic manner. One key to supporting this process is the availability of general set of geometric operators that can perform the required modeling task [27]. To operate effectively in support of domain simplifications and dimensional reductions, the geometric operators must be able to operate at a more macro-level than the topological entities typically used to key these procedures. A concept that is popular in design research that can help support these needs is a higher level representation based on features [28]. Analysis features could be any portion of the object description for which some specific idealization process will be applied. For, example the portion of a solid (Fig. 2) that is thin in one direction could be considered a shell feature idealized by its mid-surface, and the ring beam could be represented by a beam feature. Following the approach of idealization through features, the idealization processes of geometric simplification and dimensional reduction would consist of:

1. Geometrically identifying and isolating the analysis features to be idealized [29].
2. Specifying the desired idealization process for those analysis features in terms of an appropriate set of geometric operations.
3. Invoking the geometric operators needed to idealize the feature.

In addition to the definition of the domain of an idealized model, the input to an analysis procedure requires the attribute information required to specify the material constitutive relations, the loading conditions, and the boundary conditions. The analysis attribute information must also be extracted from the object's description in the design system accounting for the required idealization processes. Since the definition of the analysis attribute information must be related to the geometric definition of the domain, the interaction between the idealization procedures and the design modeling system is through a general set of geometric operators [27].

An area critical to the reliable creation of idealized analysis models is the ability to discretize the analysis domain into a valid computational mesh (Fig. 3). The ability to automatically discretize a

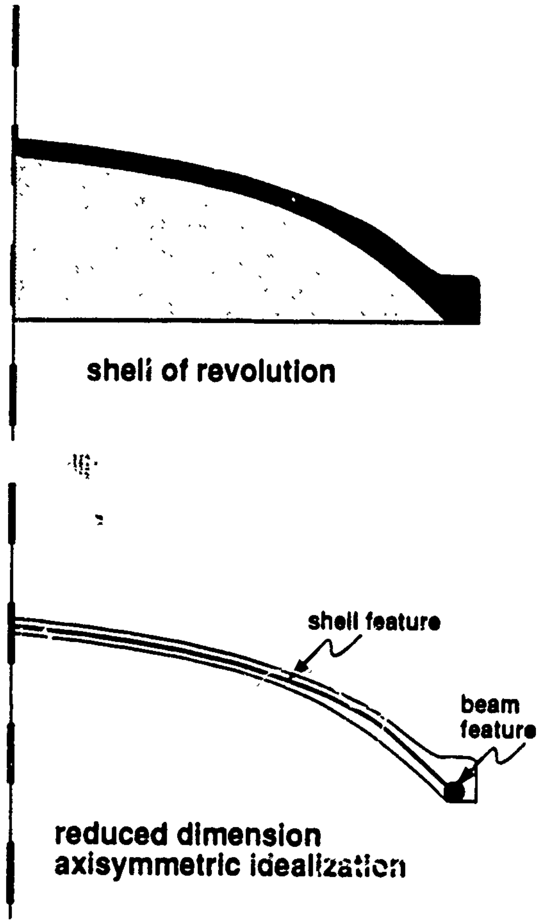


Figure 2. Shell idealization

given domain into a computational mesh is a computer-aided geometric modeling issue that has received considerable attention in the past several years [30]. One of the most difficult issues associated with the development of a general automatic mesh generation algorithm is the ability to determine if the mesh generated is a valid discretization of the domain being analyzed.

The primary source of difficulty in defining a valid mesh is that, until each mesh entity is uniquely classified with respect to a domain entity of an equal or higher geometric dimension, the geometric shape of the mesh entity is unknown and it is not possible to determine if the mesh covers the geometric domain. The determination of the classification of a mesh entity with respect to domain entities requires consideration of the geometric shape of the domain.

One approach to addressing this complexity is to carefully build the mesh entities from the lowest geometric order with respect to the domain entities of the lowest geometric order. In this manner mesh entities can be properly classified as they are generated. The topological entities of vertices, edges, faces, and regions representing 0–, 1–, 2– and 3–dimensional entities provides a convenient abstraction for discussing this process. The bottom-up approach requires the definition of mesh entities is the following order:

1. Mesh vertices at the geometric domain vertices

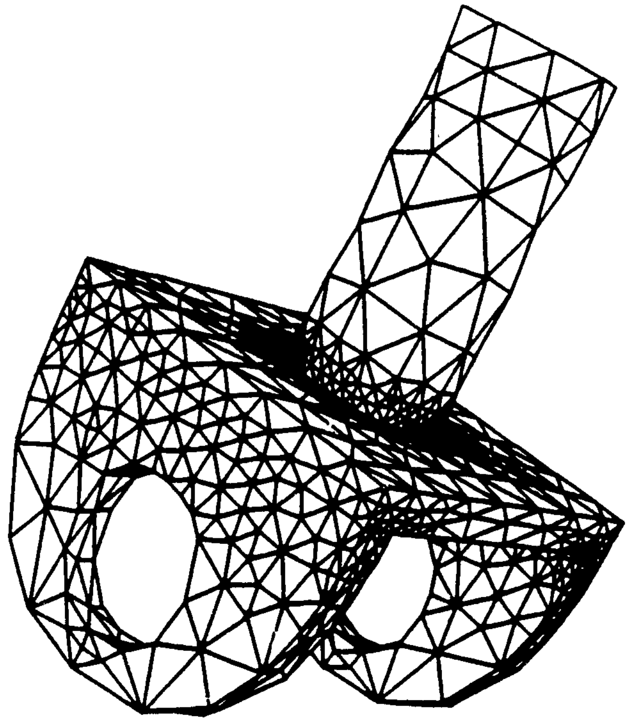


Figure 3. Automatically generated mesh

2. Mesh vertices and mesh edges on the geometric domain edges
3. Mesh vertices, mesh edges, and mesh faces on the geometric domain faces
4. Mesh vertices, mesh edges, mesh faces and mesh regions in the geometric domain regions

Each step in the process must ensure that a geometric entity is covered by the mesh and no non-existent intersections are created. One problem with this approach is that it is not consistent with the most popular algorithmic approaches to mesh generation. A less obvious, but more important issue, is that this approach is not straightforward to apply and ultimately must address the issue of determining which geometric domain entity a mesh entity is uniquely associated with.

The opposite approach to the issue of determining the validity of an automatically generated mesh is to generate a mesh without specific concern for the issue of validity and to check and correct the mesh afterwards. This approach is motivated by the popular Delaunay meshing algorithms [31], [32], [33] which generates a mesh within the convex hull defined by a set of points placed within and on the boundary of an object. In general, the resulting mesh is not a valid discretization of the domain since the mesh entities cannot be uniquely associated with the domain entities of equal or higher geometric dimension. To address this issue, the concept of a geometric triangulation has been introduced [32], [33]. A geometric triangulation represents a valid discretization of a domain into a computational mesh which meets a specific set of topological restrictions, ensures each mesh entity is uniquely classified, and the mesh is compatible with the geometry. A mesh is compatible with the geometry if all geometric entities are properly covered by mesh entities with no non-existent intersections created. For a complete technical description of a geometric triangulation as well as

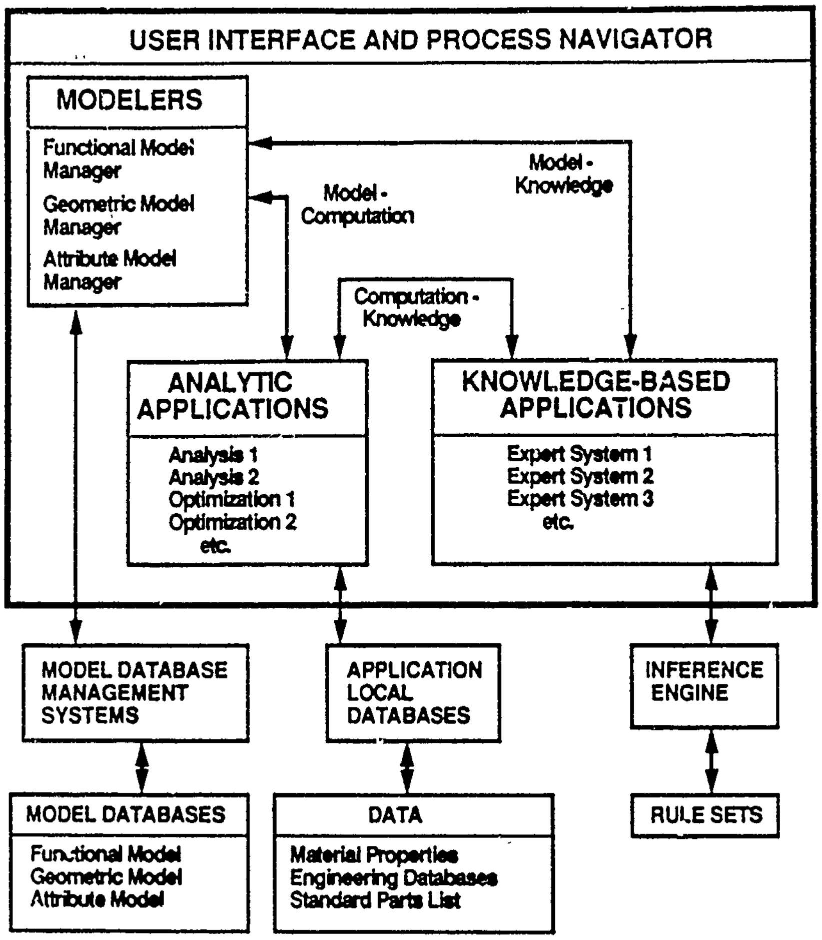


Figure 4. Modeling system supporting idealization

an algorithm for converting a Delaunay triangulation into a geometric triangulation see references [32], [33].

5. MODELING SYSTEM SUPPORTING IDEALIZATION CONTROL

The ability to apply idealization control requires a design modeling system which can house various levels of analysis idealization with smart design methodologies and engineering analysis tools. Figure 4 shows the overall framework of an engineering modeling system for mechanical objects that is specifically structured to support the idealizations used in engineering modeling and analysis [34], [35].

This system framework is consistent with the architectures being considered to support design modeling systems [36], [28], [37]. It does not represent an entire design system since it includes only the model representations and analysis tools. As indicated by Dixon, et al. [36] the analysis tools only provide the data needed for evaluation in design. However, it is important to note that the model representations determined most appropriate for supporting analysis idealization processes contain both functional and geometric models as in design systems. This level of representation is needed to support knowledge-based applications and feature representations [28].

The heart of the system is the representation of the object being designed and the modelers that support that representation. To support the functions necessary in the design evolution of an object, its representation is housed in linked functional, geometric and attribute model structures, each of which are controlled by the appropriate modelers. The other operational components of the modeling system are the applications. The applications include analysis procedures to answer performance questions, algorithms to alter the design based on analysis results, and procedures to plan the manufacturing processes, etc. Applications are separated into two groups based on the technology underlying their implementation, not on the functions addressed. The first group are analytic applications. The vast majority of the applications in this group are numerically based analysis and optimization procedures. The second group are knowledge-based applications. Knowledge-based applications are assumed here to operate from codified heuristics placed in rule sets.

The task of analysis idealization control falls to the process navigator. The process navigator interacts with the models, applications, and databases to track the various activities that have been performed and guide the application of those that are requested. By tracking the idealizations used and analyses performed to the current point in the design, the process navigator provides the designer with: 1) guidance as to the next steps in the process, 2) feedback as to the appropriateness of performing the next task request, and 3) directions to the applications appropriate to performing the requested task.

A primary goal of the process navigator for an analysis application is to provide the best balance of idealization error control possible. This includes the elimination of invalid combinations of idealizations [1], [4] and the coordination of idealization control from various sources during the engineering design process [3].

A process navigator is being implemented in terms of three components: the request interpretor, the analysis strategist and the process monitor [34],[35]. At the most basic level, the request interpretor is responsible for accepting a request to perform an operation, determining if the basic information and capabilities required to perform the task exist, and to invoking the analysis strategist to carry out the request. The analysis strategist is responsible for formulating and controlling the process steps required to perform the request. The process monitor is responsible for maintaining information about the status of the design and the tasks that have been performed previously. Given an analysis request and the current state of the design, the analysis strategist must be able to apply the various levels of idealization error control on each source of idealization error to produce the most reliable solution possible. An analysis strategist must employ feedback procedures to exercise the various levels of idealization control. The ability of the analysis strategist to interact with the design process is aided by the process monitor which maintains the appropriate information about the state of the design and the analyses performed to date. One tool critical to the proper functioning of the process monitor is the analysis goal graph [38] which interacts with the information in the design system.

6. APPROACHES TO THE CONSTRUCTION OF IDEALIZATION CONTROL PROCEDURES

The reliability of an engineering analysis process can be improved by better idealization control techniques. This section briefly discusses approaches to providing improved analysis idealization control for two common classes of mechanical analyses. In both cases the procedures proposed rely on techniques that are currently available. It is the combination of these techniques into a set of analysis idealization control procedures that provides a design engineer with the ability to more reliably perform various levels of analysis.

6.1 Analysis Idealizations Common to Airframe Modeling

The design of the structure defining the airframe is a complex process which employs a number of levels of structural analyses. The analysis goal graph and idealization control techniques required to support this design process must include [39], [3]:

1. Simple beam analysis to determine overall load distribution
2. Internal load distributions based on one-dimensional frame and truss members to determine member section property requirements
3. Overall static stress analysis representing main members as combinations of one- and two-dimensional entities employing various levels of idealization rules
4. Overall vibration analysis using appropriate sets of member stiffness and mass property idealization
5. Component fatigue and failure analyses using appropriate local idealizations and boundary conditions obtained from the previous analyses

Today the idealization processes used employ rules [40], [41], [16], [39]. These rules are based on a combination of experience, test results and analytical results, but are applied as a set of rules that are computerized through a rule base applied through an inference engine. However, as pointed out by Szabo [42] for the analysis of connections, it is important that the methods of idealization control become more rational and consistent. This requires the proper application and coordination of various levels of idealization control possible for these classes of analysis. These range from the compiled procedures [39] to adaptive procedures based on convergent sequences [21].

One current effort is focused on a prototype implementation of knowledge-based idealizations associated with the structural analysis of airframes for internal load distributions [40], [34], [35], [38]. The prototype system is being implemente primarily within the framework of the KEE environment [43] which is used to house and control the functional model (Figure 5) and structural ideaiization rules for airframes (Figure 6). The procedures are being implemented within the framework of a general process navigator so that additional idealization control procedures can be added to the system.

6.2 Geometric and Boundary Condition Simplifications in Stress Analysis

The tools needed for reliable two-dimensional stress analysis of a given mathematical problem in a cost effective manner are now available [44], [45], [23], [30]. However, their effective application in design requires the development of procedures to control the idealization errors associated with the geometric and boundary condition simplifications present during the design process (Fig. 1). These geometric and boundary condition simplifications often exist because the design has not progressed to the point where they are completely defined. However, there is the need to perform analyses at that time so that design parameters can be properly determined. The range of simple linear two-dimensional analysis requests that need to be supported during design include:

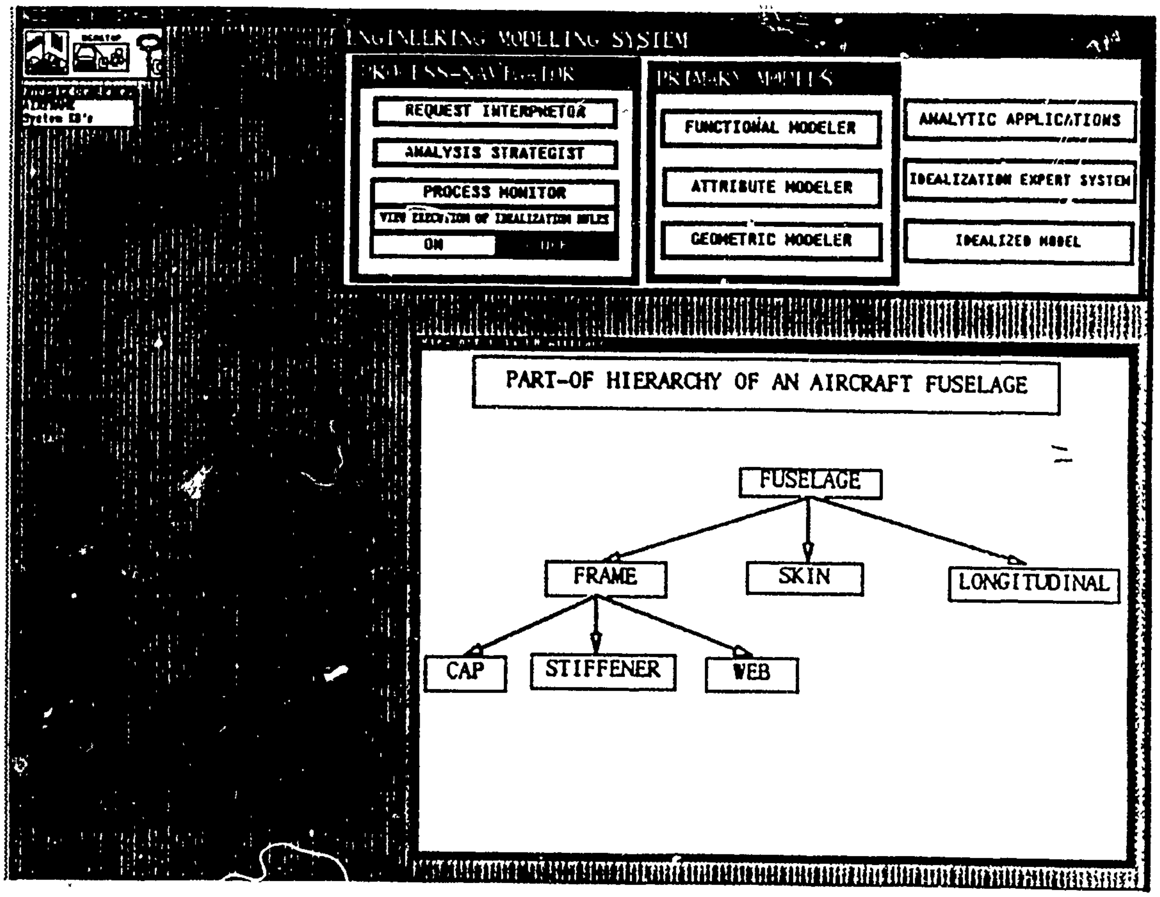


Figure 5. Functional model for portion of airframe

1. Determine load paths and load transfer of a basic configuration
2. Determine overall deformation and vibration characteristics
3. Determine all areas of likely stress concentrations
4. Evaluate the stress concentrations
5. Determine fatigue and fracture characteristics

Idealizations common to these analyses include:

1. Accounting for smooth cutouts not included in the analysis model
2. Accounting for reentrant corners that are sharp in the analysis model but will be smooth in the final design
3. Dealing with load distributions that are not yet fully defined
4. Dealing with support conditions that are not fully defined

Since these idealizations can be performed in conjunction with automated, adaptive analysis, the majority of the idealization control procedures should employ analysis feedback.

An example of such a feedback procedure has been suggested for determining when circular holes can be ignored and when they must be included in a two-dimensional stress analysis [20].

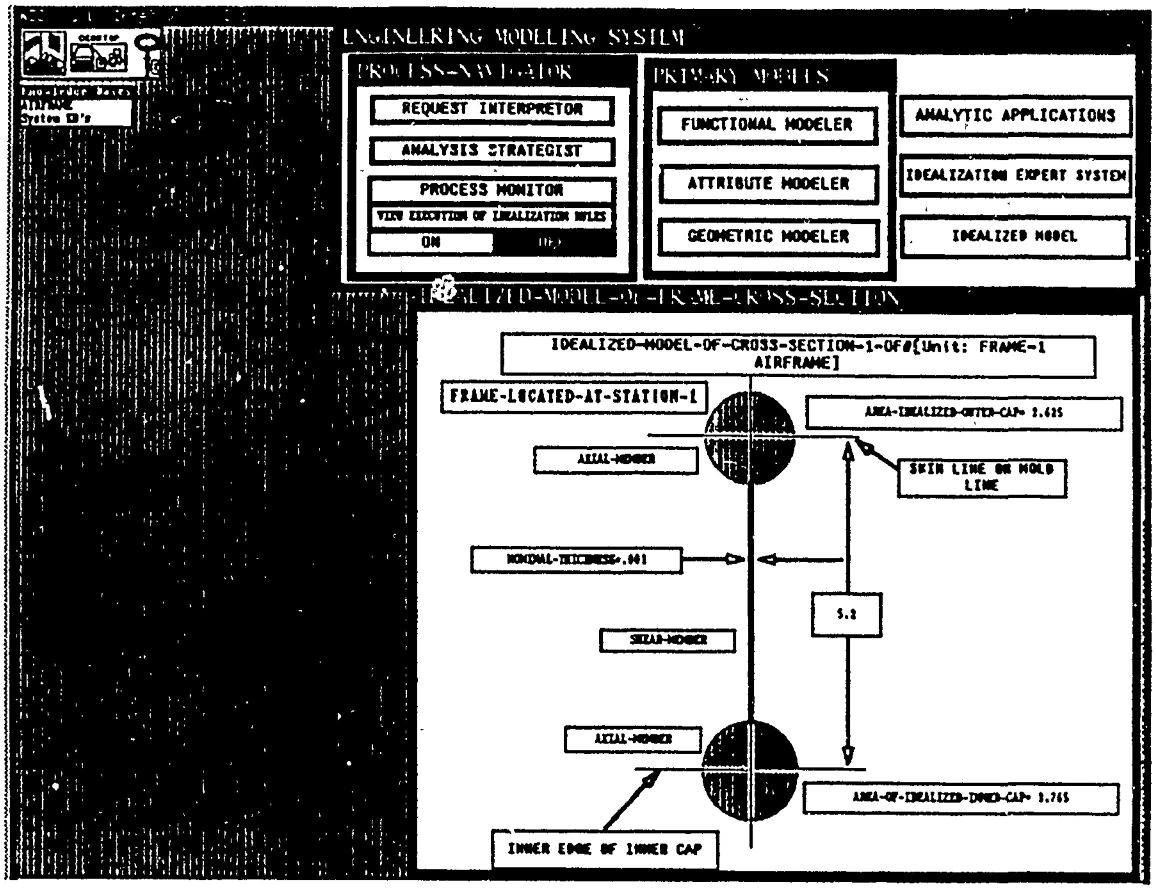


Figure 6. Dimensional reduction idealization for frame member

This procedure uses the results from an analysis in combination with analytic solutions to determine if each of the candidate holes must be included in an analysis. The first step in the process is to geometrically determine candidate holes. A candidate hole is one that is assumed to be small enough to cause only local variations in the load paths through the structure. A candidate hole must be far enough from a boundary that known analytic expressions could be used as a meaningful indicator. The second step is to perform a stress analysis with all candidates eliminated. The third step is to estimate the peak stress around each hole by multiplying the stress determined in the area of the hole by a conservative stress concentration factor derived from analytic solutions to specific relevant situations. The analytic stress concentration values are readily available for the basic situations of concern in engineering handbooks [46]. If the peak stress for the hole is below the threshold of interest to the design, that idealization is accepted; if it is above, the idealization must be improved. Improvement of the idealization could be to simply use the conservative estimate just obtained, or perform another generalized analysis including the expected critical holes in the analysis.

Consider the reverse case of a designer wishing to perform a stress analysis of a plate where there are to be lightening holes, which are not yet defined, but are known to be within specific limits of size and spacing. The process navigator must invoke an idealization procedure to reduce the thickness of the plate, based on an estimate of the volume of the holes, to represent the plate's proper overall stiffness. It would then invoke the procedure to estimate the stress concentrations that would be introduced around the holes and inform the designer if the stresses around the holes are expected to be critical.

The methods for accounting for sharp reentrant corners as well as a number of the load and support simplifications introduce the need to deal with singularities in the solution process. A simple case is a request to determine the stress concentrations before reentrant corners are filleted. In this case the process navigator should indicate to the designer that if stress concentration information is to be extracted from the analysis, a preliminary set of fillets should be designed for the reentrant corners. If the analysis is performed with the sharp corners, the adaptive analysis procedures must be modified to not spend undo effort in those areas. A more difficult situation is the case of point loads or point supports which introduce non-analytic singularities into the solution. Since it is common at early stages of design to use such point loads and supports, efforts are required to either eliminate their influence on the analysis, or to replace them with equivalent approximations that do not introduce the non-analytic singularities.

7. PERFORMANCE OF AUTOMATED ADAPTIVE SYSTEMS

The combination of automatic mesh generation and adaptive analysis techniques allows for the reliable control of the idealization errors associated with the mesh discretization for each norm for which accurate a posteriori error estimates exist. The initial acceptance of these tools by the engineering community appears to require satisfaction of two conflicting requirements. The first is that the procedures operate in conjunction with the commercial finite element analysis tools they currently use. The second is that they are computationally efficient. These two criteria tend to conflict because most current commercial finite element analysis codes solve each mesh as a separate mesh using a direct equation solver. As demonstrated below, the single largest contributor to the overall computational effort in this case is the solution of the final mesh in the adaptive process. Another difficulty in convincing many of today's practicing finite element analysts to consider such techniques is their tendency to compare the total computational effort of an automated, adaptive system to the computational effort for a single analysis. They tend not to consider the analysis model building cost or the cost of solving and comparing at least two meshes to gain some confidence in the solution accuracy.

Initial experience in the application of automated, adaptive techniques indicates the need to concentrate on the minimization of the computational effort required to solve the stiffness equations. Consideration will have to be given to the use of iterative equation solvers, which when coupled with adaptively defined meshes, yields an approximately linear computational growth rate.

To demonstrate the performance of automated adaptive analyses, consider the two-dimensional linear elasticity problem shown in Figure 7. The problem is solved by adaptive h- and hp-refinement strategies. In both procedures the mesh is adapted to control the error in the energy norm.

The adaptive portion of the automated h-refinement system consists of error estimation and local remeshing. The error is calculated from the residuals of the primary solution variable. The magnitudes of the elemental errors are used to determine the levels the elements are to be refined [44]. The local remeshing procedures in the finite quadtree mesh generator are used to automatically update the mesh to the requested refinement levels [45].

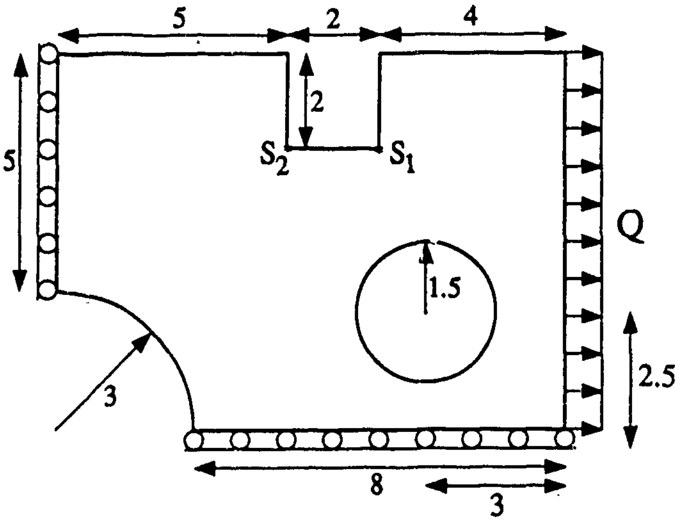


Figure 7. Problem definition

Figure 8 contains the resulting meshes consisting of quadratic elements from the h-refinement system. Shown is the original mesh with 14.1% error, the second mesh with 5.55% error, the third mesh with 2.75% error and the final mesh with 1.93% error, where the error is defined as

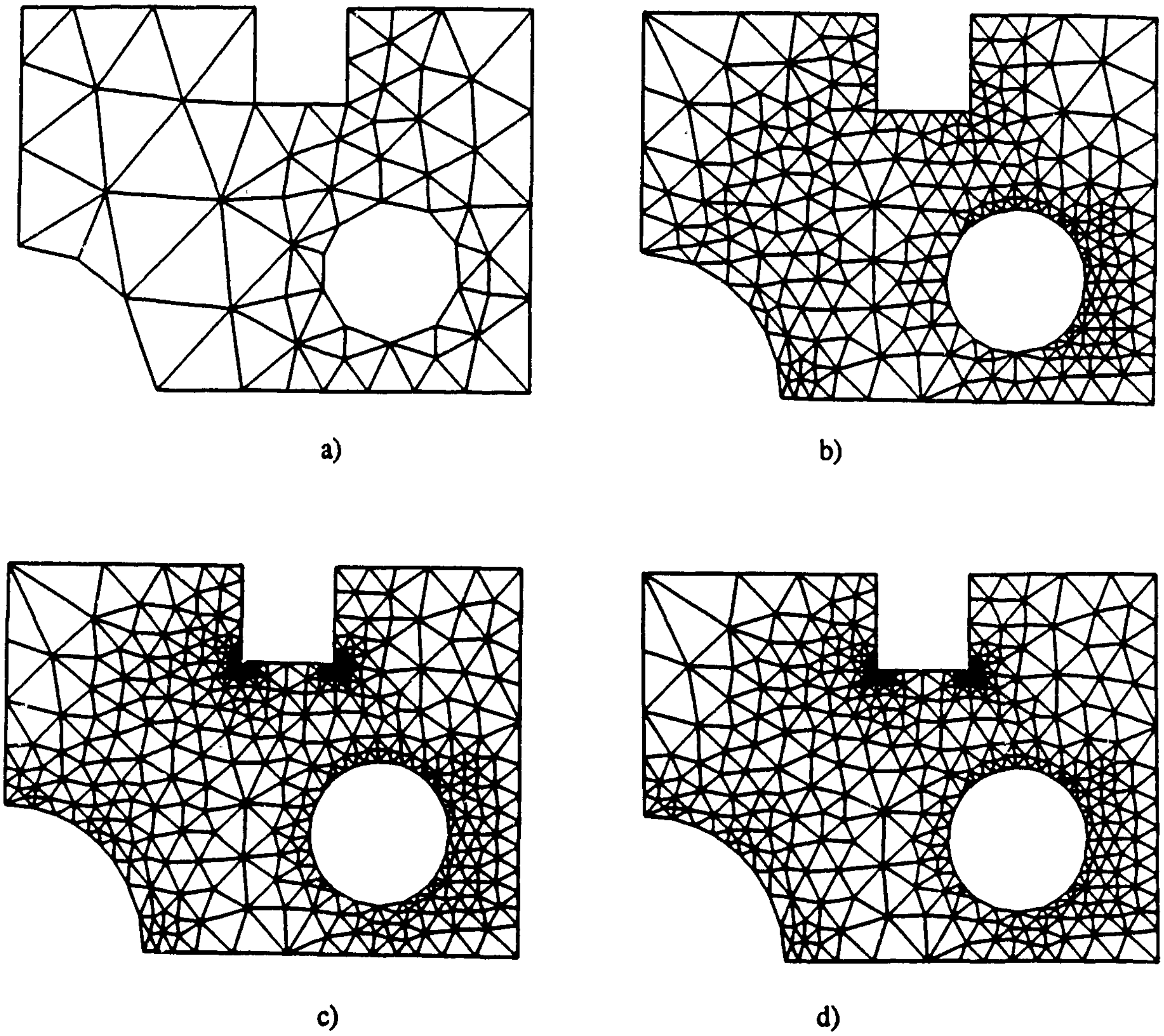
$$\eta = \left(\frac{u - U}{u} \right)^{1/2}$$

where

η = relative percent error
u = exact strain energy
U = calculated strain energy

The automated hp-refinement system employs an element removal mesh generator designed to work in conjunction with the hp-correction indication procedures in the system [23]. The hp-refinement system was defined to minimize total computation cost by using analytically-based criteria tuned by numerical experimentation[23]. To minimize the computational effort in the modeling process, error prediction techniques are used to specify the best combination finite element discretization, h, and the polynomial order, p, needed to just achieve the required accuracy. The actual solution is carried out using PROBE [47].

The starting mesh for the hp-refinement system is shown in Figure 9a. Three preliminary analyses at levels p = 2, 3, and 4 are performed to determine information that is needed by the adaptive analysis procedures. The relative error at p = 4 is η = 12.46%. Based on the feedback of the adaptive analysis procedures, the preliminary mesh is refined as shown in Figure 9b. Three additional analyses at levels p= 2, 3, and 4 are performed on the refined mesh. The relative error in the analysis of the refined mesh at p = 4 is η = 3.49%. To achieve less than 2.0% error in the energy norm, the adaptive procedures predicted the need for one additional layer of small elements around each of the two singular points, and a polynomial level of p = 5. One additional layer of small elements is created



a) b)

c) d)

Figure 8. Adaptively refined meshes at a) 14.1%, b) 5.55%, c) 2.75% and d) 1.93% error.

around both singular points in the mesh shown in Figure 9b and an analysis is performed at level p = 5. The relative error in the final analysis is $\eta = 1.28\%$.

Results of the h- and hp-refinement systems are shown in Figure 10. Included with the comparisons of the above mentioned adaptive analysis systems, are the results from 3-noded linear elements in conjunction with adaptive h-refinement [48]. Figure 10a presents the relative percent error in the energy norm versus the number of degrees of freedom, and shows that the hp-refinement system reduces the error in the mesh with less degrees of freedom than that used by the h-refinement system, and the 6-noded h-refinement meshes reduce the error faster than the 3-noded h-refinement meshes. This is to be expected.

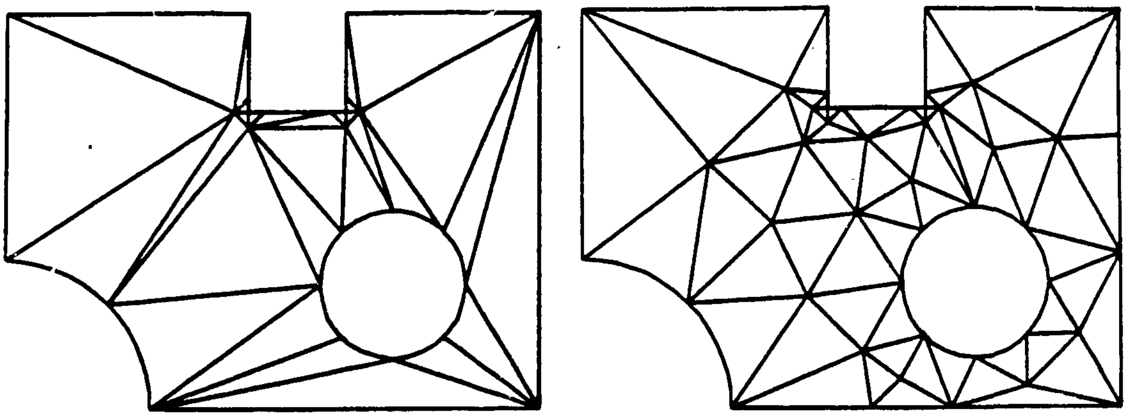


Figure 9. Adaptively refined hp meshes: a) preliminary mesh, b) final mesh

Figure 10b contains the plot of the relative percent error versus total solution time on a VAX 6320. Each point on the plot corresponds to the total computational effort required to reach that point. This includes all mesh generation, equation formulation, equation solution and error estimation efforts. In both systems the mesh generation process is under two percent of the total cost, and error estimation is under fifteen percent of the cost. The dominate cost of the process is the solution of the finest mesh. Both procedures employ direct equation solvers common to finite element analysis of linear elliptic problems (a skyline solver for the h-refinement systems and a frontal solver for the hp-refinement system). The reduction in total solution time for the hp-refinement system over the h-refinement system (Figure 10b) is greater than the reduction in numbers of degrees of freedom for a given percent error (Figure 10a). The improvement is due to the nature of direct equation solvers and the relative cost of solving the adaptive h- and hp-refinement meshes. Using the direct solvers, the time needed to solve a set of stiffness equations is proportional to the number of solution variables times the square of the average wavefront or skyline. The degrees of freedom used to define the stiffness matrices of the elements in PROBE [47] are constructed in such a manner that the average wavefront in the hp-refinement system grows in a manner similar to that experienced in the h-refinement system. Therefore, the total solution times required for adaptively defined meshes of nearly equal numbers of degrees of freedom is the same. It was also observed that on the adaptively defined meshes produced by both systems, the average wavefront growth is nearly linear with the number of degrees of freedom. Therefore, the computational advantage of the hp-refinement system over the h-refinement system is due to the combination of the following four factors:

1. Solution times for equal number of degrees of freedom nearly the same for both systems
2. Nearly linear growth of average wavefront with number of degrees of freedom for both systems
3. Solution time proportional to average wavefront squared
4. The need for a larger number of degrees of freedom to reach the same level of accuracy in the h-refinement system

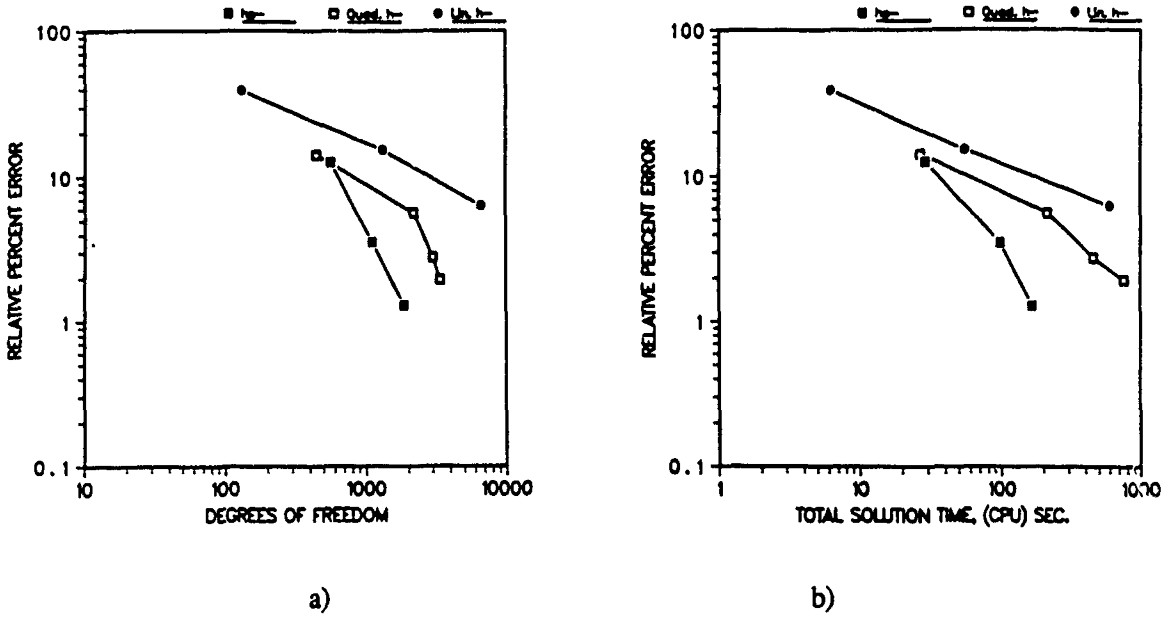


Figure 10. h- and hp-refinement results: a) relative % error versus degrees of freedom, b) relative % error versus total computational time required to reach a given relative % error.

The fact that the h-refinement procedures may use one or two additional refinement steps is not a major contributor to the difference in total computation time. Relative to the cost of the solution of the last mesh, the contribution from the solution of additional intermediate meshes is small [23].

It is important to note that the results on computational times presented here are specific to the procedures and techniques presented for the automatic solution of plane elasticity problems. Relative changes in any of the factors important to the major contributions to the computational effort will influence the relative advantage of one method over another. Based on the experience gained to date, efforts to improve procedures should concentrate on reducing the solution effort required to solve the last set of meshes.

8. CLOSING REMARKS

This paper has presented a general discussion of analysis idealization control within engineering design. An approach and framework to support the application of analysis idealization control tools was presented.

The need for design engineers to remain competitive by applying the most advanced analysis techniques available underscores the need to place an emphasis on the development of reliable idealization control techniques for those tools.

9. ACKNOWLEDGEMENT

The authors would like to acknowledge the support of the Industrial Associates sponsoring the research activities of the Program for Automated Modeling.

10. REFERENCES

[1] I. Babuska. Uncertainties in engineering design: Mathematical theory and numerical experience. In J. E. Bennett and M. E. Botkin, editors, *The Optimum Shape: Automated Structural Design*, pages 171–197, Plenum Press, NY, 1986.

[2] I. Babuska. Adaptive mathematical modeling. In J. E. Flaherty, P. J. Paslow, M. S. Shephard, and J. D. Vasilakis, editors, *Adaptive Methods for Partial Differential Equations*, pages 1–14. SIAM, 1989.

[3] M. S. Shephard. Idealization in engineering modeling and design. *Research on Engng. Design*, to appear:—, 1990.

[4] B. A. Szabo. On errors of idealization in finite element analysis of structural connections. In J. E. Flaherty, P. J. Paslow, M. S. Shephard, and J. D. Vasilakis, editors, *Adaptive Methods for Partial Differential Equations*, pages 15–28. SIAM, 1989.

[5] I. Babuska, O. C. Zienkiewicz, J. Gago, and d. A. Oliveria, editors. *Accuracy Estimates and Adaptive Refinements in Finite Element Computations*. John Wiley, Chichester, 1986.

[6] J. E. Flaherty, P. J. Paslow, M. S. Shephard, and J. D. Vasilakis, editors. *Adaptive Methods for Partial Differential Equations*. SIAM, Philadelphia, PA, 1989.

[7] I. Babuska and R. C. Morgan. Composites with a periodic structure: Mathematical analysis and numerical treatment. *Comp. and Maths. with Appls.*, 11:995–1005, 1985.

[8] R. M. Christensen. *Mechanics of Composite Materials*. John Wiley and Sons, New York, 1979.

[9] J. L. Teply and G. J. Dvorak. Bounds on the overall instanteous properties of elastic-plastic composites. *J. Mech. Phys. Solids*, 36(1):29–58, 1988.

[10] J. H. Prevost. Mechanics of continuous porous media. *Int. J. Numer. Meth. Engng.*, 18:787–800, 1980.

[11] V. C. Mow, M. K. Kwan, W. M. Lia, and M. H. Holmes. A finite deformation theory for nonlinearily permeable hydrated soft tissues. In G. W. Schmid-Schonbein, S. L. Y. Woo, and B. W. Zweifach, editors, *Frontiers in Biomechanics*, pages 153–179. Springer-Verlag, 1986.

[12] R. L. Spilker, J. K. Suh, and V. C. Mow. A finite element formulation of the nonlinear biphasic model for articular cartilage and hydrated soft tissue including strain-dependent permeability. In R. L. Spilker and B. R. Simon, editors, *Computational Methods in Bioengineering*, pages 81–92. ASME, 1988. BED - Vol. 9.

[13] A. K. Noor, W. S. Burton, and J. M. Peters. Assessment of computational models for multilayered composite cylinders. In A. K. Noor, T. Belytschko, and S. J. C., editors, *Analytical and Computational Models of Shells*, pages 419–441. ASME, CED-Vol.3, 1989.

[14] A. K. Noor and J. M. Peters. A posteriori estimates for shear correction factors in multilayered composite cylinders. *J. Engng. Mechanics Div., ASCE*, 115(6):1225–1244, 1989.

[15] J. A. Bennett and M. E. Botkin, editors. *The Optimum Shape: Automated Structural Design*. Plenum Press, New York, NY, 1986.

[16] R. Gabel, W. J. Kesack, and R. A. Reed. Planning, creating and documenting a NASTRAN finite element vibration model of a modern helicopter. NASA Contractors Report 165722, NASA Langley, Hampton, VA, 1981.

[17] B. A. Szabo. Geometric idealizations in finite element computations. *Communications in Applied Numerical Methods*, 4(3):393–400, May-June 1988.

[18] B. A. Szabo. Estimation and control of error based on p convergence. In I. Babuska, O. C. Zienkiewicz, J. Gago, and de A. Oliveria, editors, *Accuracy Estimates and Adaptive Refinements in Finite Element Computations*, pages 61–78. John Wiley, Chichester, 1986.

[19] G. Strang and G. J. Fix. *An Analysis of the Finite Element Method.* Prentice-Hall, Englewood Clifs, N.J., 1973.

[20] M. S. Shephard and M. A. Yerry. Toward automated finite element modeling. *Finite Elements in Analysis and Design*, 2:143–160, 1986.

[21] M. Vogelius. On mathematical modeling - dimensional reduction. Technical report, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland, 1980. Tech. Note BN-940.

[22] M. J. Wheeler and S. M. Yunus. An efficient error approximation technique for use with adaptive meshing. In *Proc. Second International Conference on Quality Assurance and Standards in Finite Element Analysis*, pages 211–223. NAFEMS, 1989.

[23] M. K. Georges. *A Two-Dimensional Automatic Modeling System for the h-p Version of the Finite Element Method.* PhD thesis, Rensselaer Design Research Center, Rensselaer Polytechnic Institute, Troy, NY, 1989.

[24] E. Rank and I. Babuska. An expert system for the optimal mesh design in the hp-version of the finite element method. *Int. J. Numer. Meth. Engng.*, 24:2087–2106, 1987.

[25] K. J. Weiler. The radial-edge structure: A topological representation for non-manifold geometric boundary representations. In M. J. Wozny, H. W. McLaughlin, and J. L. Encarnacao, editors, *Geometric Modeling for CAD Applications.* North Holland, 1988.

[26] K. J. Weiler. Boundary graph operators for non-manifold geometric modeling topology representation. In M. J. Wozny, H. W. McLaughlin, and J. L. Encarnacao, editors, *Geometric Modeling for CAD Applications.* North Holland, 1988.

[27] M. S. Shephard. The specification of physical attribute information for engineering analysis. *Engineering with Computers*, 4:145–155, 1988.

[28] J. R. Dixon, E. C. Libardi, and N. E. H. Unresolved research issues in development of design-with-features. In *Geometric Modeling for Product Engineering*, pages 183–196. North-Holland, Amsterdam, 1990.

[29] M. R. Henderson. Feature definition techniques analysis. In *Geometric Modeling for Product Engineering*, pages 321–335. North-Holland, Amsterdam, 1990.

[30] M. S. Shephard. Approaches to the automatic generation and control of finite element meshes. *Applied Mechanics Review*, 41(4):169–185, 1988.

[31] J. C. Cavendish, D. A. Field, and W. H. Frey. An approach to automatic three-dimensional mesh generation. *Int. J. Numer. Meth. Engng.*, 21:329–347, 1985.

[32] W. J. Schroeder. *Geometric Triangulations: with Application to Fully Automatic 3D Mesh Generation.* PhD thesis, Rensselaer Design Research Center, Rensselaer Polytechnic Institute, Troy, NY, May 1990.

[33] W. J. Schroeder and M. S. Shephard. An O(N) algorithm to automatically generate geometric triangulations satisfying the delaunay circumsphere criteria. *Eng. with Computers*, 5(3/4):177–194, 1989.

[34] E. V. Korngold, M. S. Shephard, R. Wentorf, and D. L. Spooner. Architecture of a design system for engineering idealizations. In *Advances in Design Automation - 1989: Vol. 1, Computer Aided and Computational Design*, pages 259–265. ASME, 1989.

[35] M. S. Shephard, E. V. Korngold, and R. Wentorf. Design systems supporting engineering idealizations. In *Geometric Modeling for Product Engineering*, pages 279–300. North-Holland, Amsterdam, 1990.

[36] J. R. Dixon, M. J. Guenette, R. K. Irani, E. H. Nielsen, M. F. Orelup, and R. V. Welch. Computer-based models of design processes: The evaluation of designs for redesign. In *Preprints of NSF Engineering Design Research Conference*, pages 491–506, U. Massachusetts, Amherst, MA, 1989.

[37] S. A. Safier and M. S. Fox. The role of archtiecture in computer-assisted design systems. In *Preprints of NSF Engineering Design Reaserch Conference*, pages 507–520, U. Massachusetts, Amherst, MA, 1989.

[38] R. Wentorf, M. S. Shephard, and E. V. Korngold. The use of functional models to control engineering idealizations. In *Proc. of 1989 ASME International Computers in Engineering Conference*. ASME, 1989.

[39] M. C. Y. NIU. *Airframe Structural Design*. Conmilit Press Ltd., Hong Kong, 1988.

[40] A. Budhiraja, R. Wentorf, M. S. Shephard, and E. V. Korngold. The functional model of an aircraft fuselage and idealization procedures for internal load distribution. Technical Report TR-89059, Rensselaer Polytechnic Institute, Rensselaer Design Research Center, Program for Automated Modeling, 1989.

[41] R. V. Dompka, M. C. Sciascia, D. R. Lindsay, and Y. T. Chung. Plan, formulate and discuss a nastran finite element vibrations model of the bell acap helicopter airframe. Technical report, NASA, Langley Research Center, Hampton, VA, 1989. NASA Contractor Report 181774.

[42] B. A. Szabo. Hierarchic plate and shell models based on p-extension. In A. K. Noor, T. Belytschko, and S. J. C., editors, *Analytical and Computational Models of Shells*, pages 317–331. ASME, CED-Vol.3, 1989.

[43] IntelliCorp, Inc. *KEEtutor: A Basic Course*, 1988.

[44] P. L. Baehmann. *Automated Finite Element Modeling and Simulation*. PhD thesis, Rensselaer Design Research Center, Rensselaer Polytechnic Institute, Troy, NY, May 1989.

[45] P. L. Baehmann and M. S. Shephard. Adaptive multiple level h-refinement in automated finite element analyses. *Eng. with Computers*, 5(3/4):235–247, 1989.

[46] R. J. Roark and W. C. Young. *Formulas for Stress and Strain*. McGraw-Hill, New York, NY, fifth edition, 1975.

[47] B. A. Szabo. *PROBE: Theoretical Manual*. NOETIC Technology Corporation, 7980 Clayton Road, Suite 205, St. Louis, MO 63117, 1985.

[48] M. S. Shephard, Q. Niu, and P. L. Baehmann. Some results using stress projectors for error indication and estimation. In J. E. Flaherty, P. J. Paslow, M. S. Shephard, and J. D. Vasilakis, editors, *Adaptive Methods for Partial Differential Equations*, pages 83–99. SIAM, 1989.

MIXED FINITE ELEMENT METHODS FOR ELLIPTIC PROBLEMS*

DOUGLAS N. ARNOLD†

Abstract. This paper treats the basic ideas of mixed finite element methods at an introductory level. Although the viewpoint presented is that of a mathematician, the paper is aimed at practitioners and the mathematical prerequisites are kept to a minimum. A classification of variational principles and of the corresponding weak formulations and Galerkin methods—displacement, equilibrium, and mixed—is given and illustrated through four significant examples. The advantages and disadvantages of mixed methods are discussed. The concepts of convergence, approximability, and stability and their interrelations are developed, and a résumé is given of the stability theory which governs the performance of mixed methods. The paper concludes with a survey of techniques that have been developed for the construction of stable mixed methods and numerous examples of such methods.

Key words. mixed method, finite element, variational principle

1. Introduction. The term mixed method was first used in the 1960's to describe finite element methods in which both stress and displacement fields are approximated as primary variables. We begin with the most classical example, the system of linear elasticity.

The equations of linear elasticity consist of the constitutive equation

$$A\mathcal{S} = \mathcal{E}(\mathrm{u}) \text{ in } \Omega$$

and the equilibrium equation

$$\operatorname{div}\mathcal{S} = \mathrm{f} \text{ in } \Omega.$$

Here Ω denotes the region in three dimensional space, $\mathbf{R}^3$, occupied by the elastic body, $\mathrm{u} : \Omega \rightarrow \mathbf{R}^3$ denotes the displacement field, $\mathcal{E}(\mathrm{u})$ denotes the corresponding infinitesimal strain tensor, (i.e., the symmetric part of the gradient of u, $\epsilon_{ij}(\mathrm{u}) = (u_{i,j} + u_{j,i})/2)$), f denotes the imposed volume load, and $\mathcal{S} : \Omega \rightarrow \mathbf{R}^{3\times 3}_s$ (the space of symmetric 3×3 tensors) denotes the stress field. The divergence of $\mathcal{S}$, $\operatorname{div}\mathcal{S}$, is applied to each row of $\mathcal{S}$, so that $(\operatorname{div}\mathcal{S})_i = \sum_j s_{ij,j}$. The material properties are determined by the compliance tensor A which is a positive definite symmetric operator from $\mathbf{R}^{3\times 3}_s$ to itself,[1] possibly depending on the point $\mathrm{x} \in \Omega$. The constitutive equations can equally well be written as

$$\mathcal{S} = C\mathcal{E}(\mathrm{u}) \text{ in } \Omega$$

*This work was supported by NSF grant DMS-89-02433.

†Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania 16827.

[1] This means that the action of A can be written as $(A\mathcal{S})_{ij} = \sum_{kl} a_{ijkl} s_{kl}$ with the components a_{ijkl} satisfying the usual major symmetries $a_{ijkl} = a_{klij}$, minor symmetries $a_{ijkl} = a_{jikl}$, and positivity condition $\sum_{ijkl} a_{ijkl} s_{ij} s_{kl} \geq \gamma \sum_{ij} s_{ij}^2$, for all $\mathcal{S}$, where $\gamma > 0$.

where the elasticity tensor $\mathsf{C} : \mathbb{R}^{3\times 3}_s \rightarrow \mathbb{R}^{3\times 3}_s$ is the inverse of A. This example also serves to illustrate our font conventions: vector quantities are notated in boldface, second order tensors are in script, and fourth order tensors are sans serif.

To determine a unique solution, we supplement the elasticity equations by the boundary conditions

$$\mathbf{u}|_{\Gamma_d} = \mathbf{g}_d \qquad \text{and} \qquad \mathcal{S}\mathbf{n}|_{\Gamma_t} = \mathbf{g}_t$$

where Γ_d and Γ_t are complementary parts of $\partial\Omega$ and $\mathbf{g}_d$ and $\mathbf{g}_t$ give the displacements and tractions prescribed on Γ_d and Γ_t respectively.

Mixed methods for the elasticity problem are mostly based on the following *mixed variational principle* which is a form of the Hellinger–Reissner principle:

> The solution $(\mathcal{S}, \mathbf{u})$ of the elasticity problem can be characterized as the unique critical point of the functional
>
> $$L(\mathcal{T}, \mathbf{v}) = \int_\Omega \left(\frac{1}{2}\mathsf{A}\mathcal{T} : \mathcal{T} + \operatorname{div}\mathcal{T}\cdot\mathbf{v} - \mathbf{f}\cdot\mathbf{v}\right) - \int_{\Gamma_d} \mathbf{g}_d \cdot (\mathcal{T}\mathbf{n})$$
>
> over the space of all symmetric tensorfields $\mathcal{T}$ satisfying the traction boundary condition $\mathcal{T}\mathbf{n}|_{\Gamma_t} = \mathbf{g}_t$, and all vectorfields $\mathbf{v}$.

Indeed, if we set the first variation of L with respect to $\mathcal{T}$ equal to zero, we get the equation

$$\int_\Omega (\mathsf{A}\mathcal{S} : \mathcal{T} + \operatorname{div}\mathcal{T}\cdot\mathbf{u}) = \int_{\Gamma_d} \mathbf{g}_d \cdot (\mathcal{T}\mathbf{n})$$

for all $\mathcal{T}$ for which $\mathcal{T}\mathbf{n}$ vanishes on Γ_t. Integrating by parts, we obtain the constitutive equation and displacement boundary condition. Taking the variation of L with respect to $\mathbf{v}$ leads immediately to the equilibrium equation. Note that in the form of the Hellinger–Reissner principle presented, the traction boundary condition is *essential*—it is imposed *a priori* on the space where the stress tensor is sought—while the displacement boundary condition arises *naturally* from the variational principle.

To make the variational principle precise, we must state over what space of functions $\mathcal{T}$ and $\mathbf{v}$ are to vary. The appropriate choice for $\mathcal{T}$ is the subspace of $\mathcal{H}(\operatorname{div})$ (symmetric tensorfields which are square integrable and have square integrable divergence) of fields satisfying the traction boundary condition, and for $\mathbf{v}$ the space $\mathbf{L}^2$ of all square integrable vectorfields. The reader who is uncomfortable with these function spaces need not be concerned: suffice it to say that they are chosen in a fairly natural way so that the integrals involved in the definition of L make sense.

A key point, which is characteristic of mixed variational principles, is that the pair $(\mathcal{S}, \mathbf{u})$ is *not* an extreme point of the Hellinger–Reissner functional. It is a *saddle point*. In fact

$$L(\mathcal{S}, \mathbf{v}) \leq L(\mathcal{S}, \mathbf{u}) \leq L(\mathcal{T}, \mathbf{u}) \quad .$$

for all $\mathcal{T} \in \mathcal{H}(\mathrm{div})$ satisfying $\mathcal{T}\mathrm{n}|_{\Gamma_t} = \mathrm{g}_t$ and all $\mathrm{u} \in \mathrm{L}^2$. It follows from this saddle point condition that

$$(1) \qquad \sup_{\mathrm{v}\in\mathrm{L}^2} \inf_{\substack{\mathcal{T}\in\mathcal{H}(\mathrm{div}) \\ \mathcal{T}\mathrm{n}|_{\Gamma_t}=\mathrm{g}_t}} L(\mathcal{T},\mathrm{v}) = L(\mathcal{S},\mathrm{u}) \qquad \text{and} \qquad \inf_{\substack{\mathcal{T}\in\mathcal{H}(\mathrm{div}) \\ \mathcal{T}\mathrm{n}|_{\Gamma_t}=\mathrm{g}_t}} \sup_{\mathrm{v}\in\mathrm{L}^2} L(\mathcal{T},\mathrm{v}) = L(\mathcal{S},\mathrm{u}).$$

Now, because u satisfies the constitutive equation, $\mathcal{E}(\mathrm{u})$ is square integrable. It follows (from Korn's inequality) that the gradient of u is square integrable, i.e., $\mathrm{u} \in \mathrm{H}^1$. Let us set, for any $\mathrm{v} \in \mathrm{L}^2$,

$$E(\mathrm{v}) = -\inf_{\substack{\mathcal{T}\in\mathcal{H}(\mathrm{div}) \\ \mathcal{T}\mathrm{n}|_{\Gamma_t}=\mathrm{g}_t}} L(\mathcal{T},\mathrm{v}).$$

Then $E(\mathrm{u}) = -L(\mathcal{S},\mathrm{u})$ and we have from (1) that

$$-E(\mathrm{u}) = \sup_{\mathrm{v}\in\mathrm{L}^2} -E(\mathrm{v}),$$

and, *a fortiori*,

$$-E(\mathrm{u}) = \sup_{\mathrm{v}\in\mathrm{H}^1} -E(\mathrm{v})$$

or

$$(2) \qquad E(\mathrm{u}) = \inf_{\mathrm{v}\in\mathrm{H}^1} E(\mathrm{v}),$$

i.e., the displacement field u is characterized as the minimizer of the functional E over H^1. We shall show in a moment that for any $\mathrm{v} \in \mathrm{H}^1$

$$(3) \qquad E(\mathrm{v}) = \begin{cases} \int \left(\frac{1}{2} C\,\mathcal{E}(\mathrm{v}) : \mathcal{E}(\mathrm{v}) + \mathrm{f}\cdot\mathrm{v}\right) - \int_{\Gamma_t} \mathrm{g}_t \cdot \mathrm{v}, & \text{if } \mathrm{v}|_{\Gamma_d} = \mathrm{g}_d \\ \infty, & \text{otherwise.} \end{cases}$$

This permits us to interpret (2) as the following variational priniciple, which is nothing but the usual prinicipal of minimal potential energy energy.

The displacement field u solving the elasticity problem minimizes the functional

$$\int_\Omega \left(\frac{1}{2} C\,\mathcal{E}(\mathrm{v}) : \mathcal{E}(\mathrm{v}) + \mathrm{f}\cdot\mathrm{v}\right) - \int_{\Gamma_t} \mathrm{g}_t \cdot \mathrm{v}$$

over the space of all vectorfields satsifying the displacement boundary conditions.

Thus, starting from the Hellinger–Reissner mixed principle, we have derived the standard displacement variational principle. Note that for the latter the displacement boundary condition is essential, and the traction condition natural.

To verify (3) we integrate by parts to get

$$L(\mathcal{T},v)=\int_\Omega\left(\frac{1}{2}A\mathcal{T}:\mathcal{T}-\mathcal{T}:\mathcal{E}(v)-f\cdot v\right)+\int_{\Gamma_t} g_t\cdot v+\int_{\Gamma_d}(v-g_d)\cdot(\mathcal{T}n).$$

Now if $v-g_d$ doesn't vanish on Γ_d, then we may take $\mathcal{T}$ such that $\mathcal{T}n$ is an arbitrarily large negative multiple of this quantity on Γ_d, and we can arrange as well that $\mathcal{T}$ decay quickly away from $\partial\Omega$ so that its L^2 norm is arbitrarily small. It follows that $L(\mathcal{T},v)$ can be made negative with arbitrarily large magnitude by appropriate choice of $\mathcal{T}$. Thus, if $v|_{\Gamma_d}\neq g_d$, then $E(v)=\infty$. On the other hand, if $v|_{\Gamma_d}=g_d$, then

$$L(\mathcal{T},v)=\int_\Omega\left(\frac{1}{2}A\mathcal{T}:\mathcal{T}-\mathcal{T}:\mathcal{E}(v)-f\cdot v\right)+\int_{\Gamma_t} g_t\cdot v.$$

This quantity is clearly minimal when $A\mathcal{T}=\mathcal{E}(v)$, i.e, when $\mathcal{T}=C\,\mathcal{E}(v)$, and in this case

$$L(\mathcal{T},v)=\int_\Omega\left(-\frac{1}{2}C\,\mathcal{E}(v):\mathcal{E}(v)-f\cdot v\right)+\int_{\Gamma_t} g_t\cdot v,$$

as claimed.

We have seen how the stress field can be eliminated from the mixed variational principle, leaving a variational characterization of the displacement. In a similar (simpler) way we can eliminate the displacement and obtain the following variational characterization of the stress: of all tensorfields which satisfy the equilibrium equation and the traction boundary conditions, $\mathcal{S}$ minimizes the complementary energy functional

$$E_c(\mathcal{T})=\int_\Omega\frac{1}{2}A\mathcal{T}:\mathcal{T}-\int_{\Gamma_d} g_d\cdot(\mathcal{T}n).$$

These three basic variational principles for linear elasticity are summarized in Table 1. For each of these variational principles, the critical point is determined by the vanishing of the first variation, which leads to a weakly formulated boundary value problem. The weak formulations corresponding to our three variational principles are given in Table 2.

Each of the three variational principles may be discretized by seeking a critical point of the relevant functional over a finite dimensional subspace (presumably of finite element type) of the admissable trial functions. Equivalently, in the weak formulations we can substitute the function spaces (H^1, $\mathcal{H}(\mathrm{div})$, and L^2) with finite dimensional subspaces. The resulting discretization methods are termed *Galerkin methods.* For the primal principle the resulting Galerkin methods are termed *displacement methods.* For the dual principle such methods are commonly referred to as *equilibrium methods.* For the mixed variational principle we obtain *mixed methods.* In all three cases, the determination of the discrete solution ultimately reduces to the solution of a finite system of algebraic equations.

Primal variational principle. Among all kinematically admissable vectorfields the displacement field is the unique critical point of the energy. This critical point is a minimum. I.e., $u \in H^1$, $u|_{\Gamma_d} = g_d$, and

$$E(u) = \inf_{\substack{v \in H^1 \\ v|_{\Gamma_d} = g_d}} E(v).$$

Dual variational principle. Among all statically admissable tensorfields the stress field is the unique critical point of the complementary energy. This critical point is a minimum. I.e., $\mathcal{S} \in \mathcal{H}(\mathrm{div})$, $\mathrm{div}\,\mathcal{S} = f$, $\mathcal{S}n|_{\Gamma_t} = g_t$, and

$$E_c(u) = \inf_{\substack{\mathcal{T} \in \mathcal{H}(\mathrm{div}) \\ \mathrm{div}\,\mathcal{T} = f \\ \mathcal{T}n|_{\Gamma_t} = g_t}} E_c(\mathcal{T}).$$

Mixed variational principle. Among all tensorfields assuming the prescribed tractions on Γ_t and all vectorfields, the stress and displacement fields give the unique critical point of the Hellinger–Reissner functional L. This critical point is a saddle point. I.e., $\mathcal{S} \in \mathcal{H}(\mathrm{div})$, $\mathcal{S}n|_{\Gamma_t} = g_t$, $u \in L^2$, and

$$\sup_{v \in L^2} L(\mathcal{S}, v) = L(\mathcal{S}, u) = \inf_{\substack{\mathcal{T} \in \mathcal{H}(\mathrm{div}) \\ \mathcal{T}n|_{\Gamma_t} = g_t}} L(\mathcal{T}, u)$$

TABLE 1. *Basic variational principles for linear elasticity.*

2. Other examples. Most of the elliptic problems arising from mathematical physics and engineering admit analogous variational formulations. We list some of these here. For simplicity we ignore the boundary conditions.

The scalar second order elliptic problem

$$\mathcal{A}s = \mathrm{grad}\, u, \qquad \mathrm{div}\, s = f,$$

which models, e.g., a stationary thermal distribution with temperature field u and flux field s, is entirely analogous to the linear elasticity problem.

The Kirchhoff–Love plate model may be written

$$A\mathcal{M} = -\mathcal{GRAD}\,\mathrm{grad}\, w, \qquad -\mathrm{div}\,\mathrm{div}\,\mathcal{M} = f.$$

Primal problem. Find $u \in H^1$ such that $u|_{\Gamma_d} = g_d$, and

$$\int_\Omega C\mathcal{E}(u) : \mathcal{E}(v) = -\int_\Omega f \cdot v + \int_{\Gamma_t} g_t v$$

for all $v \in H^1$ such that $v|_{\Gamma_d} = 0$.

Dual problem. Find $\mathcal{S} \in \mathcal{H}(\mathrm{div})$ such that $\mathrm{div}\,\mathcal{S} = f$, $\mathcal{S}n|_{\Gamma_t} = g_t$, and

$$\int_\Omega A\mathcal{S} : \mathcal{T} = \int_{\Gamma_d} g_d \cdot (\mathcal{T}n)$$

for all $\mathcal{T} \in \mathcal{H}(\mathrm{div})$ such that $\mathrm{div}\,\mathcal{T} = 0$ and $\mathcal{T}n|_{\Gamma_t} = 0$.

Mixed problem. Find $\mathcal{S} \in \mathcal{H}(\mathrm{div})$ satisfying $\mathcal{S}n|_{\Gamma_t} = g_t$ and $u \in L^2$ such that

$$\int_\Omega (A\mathcal{S} : \mathcal{T} + \mathrm{div}\,\mathcal{T} : u + \mathrm{div}\,\mathcal{S} : v) = \int_\Omega f \cdot v + \int_{\Gamma_d} g_d \cdot (\mathcal{T}n)$$

for all $\mathcal{T} \in \mathcal{H}(\mathrm{div})$ satisfying $\mathcal{T}n|_{\Gamma_t} = 0$ and all $v \in L^2$.

TABLE 2. *Weak formulations associated with the three variational principles.*

(Here $(\mathcal{GRAD}\,\mathrm{grad}\, w)_{ij} = w_{,ij}$ is the matrix of second partial derivatives and $\mathrm{div}\,\mathrm{div}\,\mathcal{M} = \sum_{ij} m_{ij,ij}$.) The mixed variational principle characterizes the moment tensor $\mathcal{M}$ and the transverse displacement w as a saddle point of the functional

$$\int \left(\frac{1}{2}A\mathcal{M} : \mathcal{M} + w\ \mathrm{div}\,\mathrm{div}\,\mathcal{M} + fw\right),$$

whereas the primal principle asserts that w minimizes the energy functional

$$\int \left(\frac{1}{2}C(\mathcal{GRAD}\,\mathrm{grad}\, w) : \mathcal{GRAD}\,\mathrm{grad}\, w - fw\right)$$

where $C = A^{-1}$.

In Stokes flow, the velocity u and pressure p satisfy

$$\mathrm{div}\, C\mathcal{E}(u) + \mathrm{grad}\, p = f, \qquad \mathrm{div}\, u = 0.$$

Together they are a saddle point of the the functional

$$\int \left(\frac{1}{2} C\mathcal{E}(u):\mathcal{E}(u) + p \operatorname{div} u + fu\right).$$

For Stokes problems the primal variational principle, which characterizes the pressure independently of the velocity, is rarely used. This is because it involves the inversion of the differential operator div $C\mathcal{E}(\cdot)$, which is rarely practical. On the other hand, equilibrium methods, based on the dual principle that u minimize

$$\int \left(\frac{1}{2} C\mathcal{E}(u):\mathcal{E}(u) + fu\right)$$

over divergence-free fields, are occasionally used.

The mixed weak formulations for all four examples are listed in Table 3. For simplicity we continue to ignore boundary conditions, and also do not insist on the precise function spaces involved. Notice the characteristic form shared by all the examples.

3. Advantages and disadvantages of mixed methods. A number of reasons have been put forth to prefer mixed methods over displacement or equilibrium methods in some situations. First of all, equilibrium methods are rarely used in practical computation due to the difficulty of creating finite element spaces incorporating the necessary constraints (the conditions of static admissability and, in particular, the equilibrium condition in the case of elasticity). Thus the practical choice is usually between the primal-based displacement methods and the mixed methods. For some problems, such as the Stokes problem, primal-based methods are impractical. For such problems the mixed methods are the simplest and most direct alternative and are widely used.

For the other examples, however, the most basic methods are primal-based. A commonly stated reason to prefer mixed methods in these cases is that the dual variable (stress for elasticity, flux for thermal problems, moments for plate bending) is often the variable of most interest. For primal-based methods this variable is not a fundamental unknown and must be obtained *a posteriori* by differentiation, which entails a loss of accuracy. For the mixed methods, however, the dual variable is computed as a fundamental unknown. Of course, this argument is only heuristic. Its correctness depends on the available mixed finite element spaces and primal finite element spaces, the accuracy they offer, and the computational work they require to solve the corresponding discrete problems.

Another common motivation for the use of mixed methods is the avoidance of C^1 elements for plate bending and other fourth order problems. This is because the mixed functional for plate bending involves no more than two derivatives in any term and hence, after a suitable integration by parts, may be evaluated on finite element spaces with merely continuous elements. The primal variational functional, however, requires the use of C^1 elements (or non-conforming elements).

Elasticity: Find $\mathcal{S}$ and $\mathbf{u}$ such that $\int_\Omega (A\mathcal{S}:\mathcal{T} + \operatorname{div}\mathcal{T}\cdot\mathbf{u} + \operatorname{div}\mathcal{S}\cdot\mathbf{v}) = \int_\Omega \mathbf{f}\cdot\mathbf{v}$ for all $\mathcal{T}$ and $\mathbf{v}$.
Scalar second order: Find s and u such that $\int_\Omega (\mathcal{A}\mathsf{s}:\mathsf{t} + \operatorname{div}\mathsf{t}\cdot u + \operatorname{div}\mathsf{s}\cdot v) = \int_\Omega f\cdot v$ for all t and v.
Plate: Find $\mathcal{M}$ and w such that $\int_\Omega (A\mathcal{M}:\mathcal{N} + \operatorname{div}\operatorname{div}:\mathcal{N}\, w + \operatorname{div}\operatorname{div}:\mathcal{M}\, v) = \int_\Omega fv$ for all $\mathcal{N}$ and v.
Stokes: Find $\mathbf{u}$ and p such that $\int_\Omega [C\mathcal{E}(\mathbf{u}):\mathcal{E}(\mathbf{v}) + \operatorname{div}\mathbf{v}\, p + \operatorname{div}\mathbf{u}\, q] = \int_\Omega \mathbf{f}\cdot\mathbf{v}$ for all $\mathbf{v}$ and q.

TABLE 3. *Mixed weak formulations for various problems.*

Another advantage of mixed variational principles is their robustness in the presence of certain limiting and extreme situations. For example, in the case of linear elasticity, the compliance tensor A becomes singular in the limit of incompressibility. Consequently its inverse, the elasticity tensor C, blows up: it is very large for nearly incompressible materials and infinite for incompressible ones. (E.g., in the istropic case

$$A\mathcal{T} = \frac{1+\nu}{E}\left(\mathcal{T} - \frac{\nu}{1+\nu}\operatorname{tr}(\mathcal{T})\mathcal{I}\right), \qquad C\mathcal{T} = \frac{E}{1+\nu}\left(\mathcal{T} + \frac{\nu}{1-2\nu}\operatorname{tr}(\mathcal{T})\mathcal{I}\right),$$

with $\mathcal{I}$ denoting the identity matrix and tr the matrix trace operator. As the Poisson ratio $\nu \uparrow 1/2$, A tends nicely to the limiting value

$$A\mathcal{T} = \frac{3}{2E}\left(\mathcal{T} - \frac{1}{3}\operatorname{tr}(\mathcal{T})\mathcal{I}\right),$$

but C blows up.) A analogous situation holds for the Reissner–Mindlin plate model where the robustness is with respect to the plate thickness. Robustness properties of mixed methods have also been reported in other situations, as well. That is, mixed methods have been observed to perform significantly better than closely related displacement methods in particular applications that involve some extreme or limiting behavior. For example, Ewing, Wheeler, and others have reported superior computations of pressure (which satisfies a scalar second order elliptic problem arising from Darcy's law) via mixed methods when simulating the miscible displacement of oil from a porous media [12], [13]. Marini and Savini [20] have reported improved results in semiconductor device modelling through the use of mixed methods. In each case, the mixed methods seem to be exhibiting greater robustness with respect to the roughness of the coefficients of the equations. (In both problems there is a sharply defined front across which the coefficients change rapidly.)

In addition to the situations in which mixed methods are used explicitly, there are a number of methods which have been proposed in the literature which, while resembling displacement methods, can be shown to be equivalent to mixed methods. Such methods are called *generalized displacement methods* since they lead to discrete systems involving only degrees of freedom associated to the primal variable. However the discrete system differs from what would be obtained by straightforward discretization of the primal variational principle. The best known examples are the reduced and selective integration methods in which all or some of the terms of the primal energy functional are intentionally integrated with low accuracy. This apparently paradoxical procedure of *reducing* the integration accuracy in order to *increase* the solution accuracy was poorly understood and quite controversial when first introduced. In almost every case where it is successful, however, it can be shown that such a method is equivalent to a rather natural mixed method with exact, or at least accurate, integration [19]. In a number of cases the theory of mixed methods can be applied to provide a complete understanding and justification of reduced integration procedures. (Cf. [1] where the theory of mixed methods is used to give a complete analysis of reduced integration and standard displacement methods for the Timoshenko beam problem and [18] where some reduced integration methods for the Stokes problem and Reissner-Mindlin plate are analyzed as mixed methods.) A similar situation holds for other generalized displacement methods, such as ones involving harmonic averaging of rough coefficients [6] and interpolation [8] in the computation of the stiffness matrix. In addition a number of non-conforming displacement methods can be viewed, and best analyzed, as mixed methods [2]. In our view, this constitutes one of the most important roles of the theory of mixed methods: it provides tools to design and analyze high performance generalized displacement methods.

There are also obvious disadvantages to mixed methods in comparison with displacement methods. Because both the primal and dual variable are approximated simultaneously, the discrete system will typically involve many more degrees of freedom than a displacement method which uses a similar space to approximate the primal variable (but does not directly approximate the dual variable). Morever the fact that the primal variational principle is an extremal principle is reflected as positivity of the discrete system. Thus displacement methods for all the problems we have considered lead to positive definite algebraic systems. Since the mixed variational principal is a saddle point principal rather than an extremal principal the discrete system will be indefinite, possessing both positive and negative eigenvalues. Consequently a number of solution methods, both direct methods such as Cholesky decomposition and iterative methods like conjugate gradients, can not be applied directly.

Both these objection can often be overcome in practice by implementing mixed methods as generalized displacement methods. A simple case is when the finite element space for the dual variable does not incorporate any interelement continuity, i.e., all the degrees of freedom associated with the dual variable are internal to the elements. In this case the dual variable can be eliminated at negligible cost (by static condensation). The resulting system involves only the primal degrees of freedom and is positive definite. In fact many reduced integration methods arise in this way. More generally, when all the degrees of freedom of the dual variable are either interior to elements or lie on element edges (in two dimensions) or faces (in three dimensions)—but not at vertices—there is a quite general procedure to eliminate them at little cost [2], [15]. In contrast to the completely discontinuous case, this procedure adds additional degrees of freedom for the primal variable. The generalized displacement methods which arise typically use nonconforming elements.

A third possible objection to mixed methods is that they are subject to possible instabilities which do not arise for standard displacement methods. Thus the finite element spaces used to discretize extremal variational principles may be selected considering only their approximation properties and convenience of implementation. However for mixed variational principles, when spaces are selected on this basis alone they will almost always give poor results. For good convergence, the spaces must also satisfy some rather subtle stability conditions. Consequently the theory of mixed methods is more involved (and more interesting) than for displacement methods, and the design of effective mixed methods requires more expertise than for displacement methods. The stability properties of mixed methods, which form the heart of their mathematical theory, will be the subject of the remainder of this paper.

4. Approximability, stability, and convergence of Galerkin methods. All the weak formulations we have considered—primal, dual, and mixed—can be written in the form

(4) Find $u \in V$ such that $B(u,v) = F(v)$ for all $v \in V$,

where V is some function space, $B : V \times V \to \mathbf{R}$ is a bilinear form, and $F : V \to \mathbf{R}$ is a linear form.* Indeed any linear problem arising from a variational principle (i.e., any problem

*If the problem involves inhomogeneous (i.e., nonzero) essential boundary conditions, then u here in which the solution is characterized as a critical point of some quadratic functional) has this form (although the weak form is more general—it applies to problems that don't have a variational principle). To solve such a problem by a Galerkin method, we choose a finite-dimensional subspace V_h of V (typically a space spanned by a convenient set of finite element shape functions), and determine the approximate solution u_h by the the same weak formulation, except that both the trial space where the approximate solution is sought and the space of test functions over which v varies are replaced by V_h:

$$\text{Find } u_h \in V_h \text{ such that } B(u,v) = F(v) \text{ for all } v \in V_h.$$

If the weak formulation arises from a variational principle, as in our examples, this is equivalent to discretizing the variational principle by seeking a critical point in the subspace V_h.

We shall be concerned with three properties of such Galerkin discretizations. *Convergence* measures the the smallness of the error $u - u_h$ between the exact solution and discrete solution. Good convergence properties are the fundamental goal of any numerical method. *Approximability* measures the error in the *best* approximation of u by elements of V_h, i.e., the smallest possible error between the exact solution u and *any* element of the discretization space V_h. Note that approximability depends on the choice of the space V_h and the exact solution u, but not on the particular problem under consideration. The convergence achieved by a method is clearly limited by the approximability of the subspace, but good approximability does *not* guarantee good convergence. The missing ingredient turns out to be *stability*, which refers to the continuity of the mapping from the data F to the discrete solution u_h.

To quantify these notions it is necessary to introduce norms to measure differences between functions. Let $\|v\|$ denote a norm on functions $v \in V$ (this simply means that $\|v\|$ is positive for any nonzero v, and that triangle inequality and the homogeneity condition $\|cv\| = |c|\|v\|$ hold). We will always assume that the norm is chosen so that the bilinear form B is bounded, i.e., that there is a constant K such that

$$B(v,w) \leq K\|v\|\|w\| \tag{5}$$

for all v and w in V. It is usually straightforward in practice to choose a norm so that (5) holds with K not unreasonably large. For example if

$$B(v,w) = \int a \operatorname{grad} v \cdot \operatorname{grad} w + \mathfrak{b} \cdot \operatorname{grad} v \, w + c\, v\, w,$$

represents not the solution but rather the difference between the solution and some other, arbitrarily chosen but fixed, function satisfying the essential boundary conditions. To avoid this technical complication, which is not relevant here, we shall henceforth assume that any essential boundary conditions are homogeneous.

the natural choice is

$$\|v\| = \sqrt{\int |\operatorname{grad} v|^2 + |v|^2},$$

which is the Sobolev H^1 norm, and then K would depend in a simple way on bounds for the coefficients a, b, and c. For the form

$$B\big((\mathsf{s},u),(\mathsf{t},v)\big) = \int \mathcal{A}\mathsf{s}:\mathsf{t} + \operatorname{div}\mathsf{t}\cdot u + \operatorname{div}\mathsf{s}\cdot v$$

the natural choice of norm is

$$\|(\mathsf{s},u)\| = \sqrt{\int |\operatorname{div}\mathsf{s}|^2 + |\mathsf{s}|^2 + |u|^2}.$$

This is the $H(\text{div})$ norm on s and the L^2 norm on u. The norms which arise naturally in this way are usually of practical significance. E.g., for displacement methods for elasticity, the natural norm is the energy norm. Of course we may be interested in the convergence of our method in other norms than the natural one (for example, we may be interested in the maximum of the stress rather than its root mean square). Convergence analysis in other norms than the natural one is possible, but involves further complications, and it is usually necessary to understand the convergence in the "natural norm" as a first step. In this paper we shall only consider convergence in the natural norm for the problem.

Having introduced a norm on V it is clear how to measure convergence and approximability, namely by the quantities

$$\|u - u_h\| \qquad \text{and} \qquad \inf_{v\in V_h} \|u - v\|$$

respectively. To quantify the notion of stability we must also have a norm on the space of functionals on V. For this purpose we use the *dual norm* defined by

$$\|F\|_* = \sup_{0\neq v\in V} \frac{\|Fv\|}{\|v\|}$$

for $F: V \to \mathbb{R}$. Then the *stability constant* for the Galerkin method is given by

$$C_h = \sup_{F:V\to\mathbb{R}} \frac{\|u_h\|}{\|F\|_*}.$$

That is, for any data F we consider the solution u_h to the discrete problem, and measure the size of u_h compared to the size of F. The largest value this ratio achieves, for any possible data F, is the stability constant. If we think of the discrete problem as a matrix equation, then the stability constant is just the norm of the inverse matrix. With this notation, we can state the fundamental relation between convergence, approximation, and stability:

$$\boxed{\|u - u_h\| \le K C_h \inf_{v \in V_h} \|u - v\|}$$

Since the constant K will generally not be large if the norm is chosen reasonably, this relation says that if the stability constant C_h is not large, then the error in the Galerkin solution will not be much larger than the error in the best approximation. To clarify this further, consider a sequence of subspaces V_h parametrized by a positive number h tending to zero (which could, for example, represent the mesh size, as in the standard finite element method, or the the inverse of the polynomial degree, for the p-version of the finite element method and spectral methods). Suppose that the spaces become more and more accurate, in the sense that

$$\lim_{h \to 0} \inf_{v \in V_h} \|u - v\| = 0.$$

Then if the stability constant C_h remains bounded as $h \to 0$, it follows that u_h converges to u at the same rate as the best approximation. If, on the other hand, $C_h \to +\infty$ quickly enough, in general u_h will not converge to u at all as $h \to 0$. If $C_h \to +\infty$ slowly, then u_h may still converge to u, but generally at a slower rate than the best approximation. In the case where C_h stays bounded we say that our method is *stable* (here method refers to the whole sequence of V_h, i.e., includes the mesh refinement or degree enhancement procedure). In summary, *if the method is stable, then the approximate solution converges to the exact solution at the same rate as the best approximation error.*

Remark. This basic result can be extended in two ways to cover the majority of linear finite element applications. (Many further extensions are possible as well, including to nonlinear problems.) First, we have only considered Galerkin methods where the trial space (in which u_h is sought) and the test space (over which the test function v varies) are the same space V_h. The standard mixed methods are of this sort. However, the fundamental error bound above applies equally well to the case of *Petrov–Galerkin methods* where different spaces are used. Second, we have only considered conforming methods in which the discrete problem is to find $u_h \in V_h$ such that $B(u, v) = F(v)$ for all $v \in V_h$, with the space $V_h \subset V$. If $V_h \not\subset V$ or if we use an approximate bilinear form B_h or an approximate linear form F_h on the discrete level which is unequal to the corresponding exact form B or F (e.g., because of numerical quadrature), then the method is *nonconforming*. For nonconforming methods the discrete equations

$$B_h(u_h, v) = F_h(v) \qquad \text{for all } v \in V_h$$

will in general not be satisfied by the exact solution u. The degree to which the exact solution fails to satisfy the discrete equations is called the *consistency error*. If the consistency error is appropriately quantified, the fundamental principle above extends to non-conforming methods as follows: if the method is stable, then the error in the Galerkin solution is bounded by a multiple of the sum of the approximation error and the consistency error. In this paper we will continue only to consider conforming methods, for which the consistency error is zero.

5. Stability of mixed methods. The basic theory sketched in the last section applies equally well to displacement methods and mixed methods. For example, consider again the elasticity problem, and for simplicity suppose that the boundary conditions are for vanishing displacement on the whole boundary ($\Gamma_d = \partial\Omega$, $g_d = 0$). For the primal formulation the bilinear form B in (4) is then

$$B(u,v) = \int_\Omega C\,\mathcal{E}(u):\mathcal{E}(v) \qquad \text{for } u,v \in \mathring{H}^1, \tag{6}$$

($\mathring{H}^1$ is the subspace of H^1 of functions vanishing on the boundary), while for the mixed formulation of this problem

$$B\big((\mathcal{S},u),(\mathcal{T},v)\big) = \int_O (A\mathcal{S}:\mathcal{T} + \operatorname{div}\mathcal{T}:u + \operatorname{div}\mathcal{S}:v) \text{ for } (\mathcal{S},u),(\mathcal{T},v) \in \mathcal{H}(\operatorname{div}) \times L^2$$

(cf. Table 2). A major difference between the two cases arises when we try to find finite element spaces which yield stable approximations. For displacement methods there is no difficulty. In fact *any* choice of subspaces $V_h \subset \mathring{H}^1$ yields stable approximation. This is because the bilinear form (6) is *coercive*, that is, the inequality

$$B(v,v) \geq \alpha\|v\|^2 \qquad \text{for all } v \in V$$

holds for some positive constant α. (This is ensured by Korn's inequality, which asserts the existence of such a constant α depending only on the domain Ω. In fact the primal formulations for all our examples are coercive.) Now the discrete solution $u_h \in V_h$ is defined by the equations $B(u_h,v) = F(v)$ for $v \in V_h$. Setting $v = u_h$ and invoking coercivity and the definition of the dual norm $\|\cdot\|_*$, we get

$$\alpha\|u_h\|^2 \leq B(u_h,u_h) = F(u_h) \leq \|F\|_*\|u_h\|$$

whence

$$\|u_h\| \leq \alpha^{-1}\|F\|_*.$$

Thus the stability constant C_h for this discretization is bounded by $1/\alpha$ no matter how the subspace V_h is chosen. Consequently the error will be of the same order as the error in best approximation. The choice of subspace need therefore only be guided by considerations of approximability and efficiency of implementation. In short, *Galerkin methods based on coercive formulations are always stable.**

*Here we use the fact that the test and trial spaces are identical. Petrov–Galerkin methods based on coercive formulations are not necessarily stable.

The situation for mixed methods is altogether different. For mixed methods the space V decomposes as the product of two spaces $V = S \times W$ and B has the special form

$$B((s,u),(t,v)) = a(s,t) + b(t,u) + b(s,v) \tag{7}$$

with $a : S \times S \to \mathbf{R}$ and $b : S \times W \to \mathbf{R}$ bilinear. One consequence is that for mixed formulations the bilinear form is *never* coercive and stability is by no means automatic. In fact elements which are chosen without due regard to stability will usually prove to be unstable. For example, from a naive point of view the simplest, most appealing element for the Stokes problem is the linear velocity–constant pressure element shown in Figure 1a. However the stability constant for this element is ∞ and the resulting discrete system of equations is singular on most meshes. This element is completely useless.

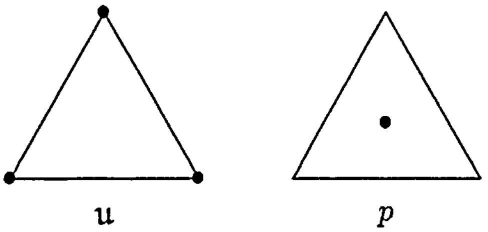


FIG. 1A. *An unstable Stokes element.*

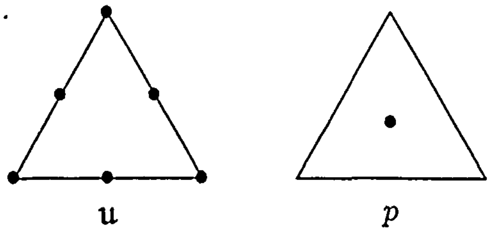


FIG. 1B. *A stable Stokes element.*

For the Stokes problem the bilinear form B takes the form (7) with

$$a(u,v) = \int_\Omega C\mathcal{E}(u) : \mathcal{E}(v) \qquad \text{for } u, v \in \mathring{H}^1,$$
$$b(q,v) = \int_\Omega q \operatorname{div} v \qquad \text{for } q \in L^2,\ v \in \mathring{H}^1.$$

Note that in this case that, although B is not coercive, at least a is. In this case it can be shown that the stability constant may be bounded in terms of the reciprocal of the coercivity constant α for a and the reciprocal of the quantity

$$\beta_h = \inf_{v \in W_h} \sup_{s \in S_h} \frac{b(s,v)}{\|s\|\|v\|}. \tag{8}$$

In particular if we choose a sequence of S_h and W_h for which β_h stays bounded away from zero, the corresponding method is stable. The condition that β_h stay bounded above zero is known as the second Brezzi condition or LBB condition.

Notice that as the space S_h increases, for fixed W_h, the constant β_h increases. In other words, for the Stokes problems (or, more generally, problems for which the a form is coercive), enrichment of the space S_h increases stability. For example the quadratic velocity–constant pressure Stokes element, shown in Figure 1b, is stable [14].

However the condition that the a form be coercive is not satisfied for most mixed methods. In fact, of the four mixed formulations presented, only that for the Stokes problem has this property. For the elasticity problem, for example, we have $a(\mathcal{T},\mathcal{T}) = \int_\Omega A\mathcal{T} : \mathcal{T}$. Since it is possible to find $\mathcal{T}$ which is bounded by 1 everywhere but for which the divergence of $\mathcal{T}$ is arbitrarily large, there cannot exist a constant α such that $a(\mathcal{T},\mathcal{T}) \geq \|\mathcal{T}\|^2_{\mathcal{H}(\mathrm{div})}$ for all $\mathcal{T}$ in $\mathcal{H}(\mathrm{div})$. So the a form is indeed not coercive.

However, it turns out that one can get by with a weaker condition than coercivity of a on S, namely coercivity on a particular subspace of S_h. More precisely, suppose that there exists a positive constant α_h such that

$$(9) \qquad a(z,z) \geq \alpha_h \|z\|^2 \qquad \text{for all } z \in Z_h$$

where

$$Z_h = \{\, z \in S_h \,|\, b(z,v) = 0 \text{ for all } v \in W_h \,\}.$$

Then the stability constant may be bounded in terms of the reciprocals of the constants α_h in (9) and β_h in (8). Thus if for a sequence of subspaces $S_h \times W_h$ the α_h remain bounded uniformly above zero (this is the first Brezzi condition), and the β_h do likewise (second Brezzi condition), then the resulting method is stable. This is the content of Brezzi's theorem [10].

Let us briefly indicate the idea behind the theorem. Stability refers to the invertibility of the matrix representing the discrete problem, and the stability constant is the norm of the inverse matrix. For mixed methods, the matrix has the form

$$(10) \qquad \begin{pmatrix} \mathcal{A} & \mathcal{B}^t \\ \mathcal{B} & 0 \end{pmatrix} : \begin{Bmatrix} S_h \\ W_h \end{Bmatrix} \rightarrow \begin{Bmatrix} S_h \\ W_h \end{Bmatrix}.$$

The space Z_h introduced above is the nullspace of the $\mathcal{B}$. Therefore, if we partition S_h as $Z_h \times Z_h^\perp$, where $Z_h^\perp$ denotes the orthogonal complement of Z_h in S_h, then the action of $\mathcal{B}$ on S_h may be written as

$$\begin{pmatrix} 0 & \bar{\mathcal{B}} \end{pmatrix} : \begin{Bmatrix} Z_h \\ Z_h^\perp \end{Bmatrix} \rightarrow W_h$$

where $\bar{\mathcal{B}}$ denotes the restriction of $\mathcal{B}$ to $Z_h^\perp$. Now the second Brezzi condition just asserts the invertibility of $\bar{\mathcal{B}}$. Similarly, let us decompose the action of $\mathcal{A}$ as, say,

$$\begin{pmatrix} \bar{\mathcal{A}} & \mathcal{Q} \\ \mathcal{R} & \mathcal{S} \end{pmatrix} : \begin{Bmatrix} Z_h \\ Z_h^\perp \end{Bmatrix} \rightarrow \begin{Bmatrix} Z_h \\ Z_h^\perp \end{Bmatrix}$$

Thus $\bar{\mathcal{A}}$ is the matrix associated with the bilinear form a restricted to $Z_h \times Z_h$, and the first Brezzi condition simply asserts the invertibility of this operator. The whole matrix (10), rewritten in terms of these new notations, is

$$\begin{pmatrix} \bar{\mathcal{A}} & \mathcal{Q} & 0 \\ \mathcal{R} & \mathcal{S} & \bar{\mathcal{B}}^t \\ 0 & \bar{\mathcal{B}} & 0 \end{pmatrix} : \begin{Bmatrix} Z_h \\ Z_h^\perp \\ W_h \end{Bmatrix} \rightarrow \begin{Bmatrix} Z_h \\ Z_h^\perp \\ W_h \end{Bmatrix},$$

or, rearranging rows and columns,

$$\begin{pmatrix} \bar{\mathcal{B}}^t & \mathcal{R} & \mathcal{S} \\ 0 & \bar{\mathcal{A}} & \mathcal{Q} \\ 0 & 0 & \bar{\mathcal{B}} \end{pmatrix} : \begin{Bmatrix} W_h \\ Z_h \\ Z_h^\perp \end{Bmatrix} \to \begin{Bmatrix} Z_h^\perp \\ Z_h \\ W_h \end{Bmatrix}.$$

From the upper triangular form, it is clear that the invertibility of $\bar{\mathcal{B}}$ (which is equivalent to the invertibility of $\bar{\mathcal{B}}^t$) and the invertibility of $\bar{\mathcal{A}}$ are together are necessary and sufficient for the invertibility of the whole matrix.

While Brezzi's theorem furnishes us with relatively concrete conditions which yield stability, the verification of these conditions can be quite difficult. A number of analytic techniques have been developed that ease the task some what, for example, localization theorems [9], the use special mesh-dependent norms [7], etc. We shall not go into any of these techniques here, but in the next section we discuss a number of elements that have, in one way or another, been shown to be stable.

6. The construction of stable mixed elements. In § 4 we saw that the accuracy of a finite element discretization is determined by the approximability of the exact solution by the finite element subspace and the stability of the discretization. These two properties, together with implementational issues, furnish the major factors for the construction and evaluation of the finite element spaces to be used. In § 5 we saw that stability is automatic for coercive methods, such as most displacement methods, so that the finite element space can be chosen on the basis of approximation and ease of implementation alone. However, for mixed methods the question of stability is paramount.

Various techniques have been developed for the design of stable mixed elements. In this section we review some of these techniques and some of the resulting elements. We emphasize that this review is by no means exhaustive, neither with regard to the techniques nor to the resulting methods.

As remarked above, for the Stokes problem, in which the a form is coercive, stability can always be achieved by adequate enrichment of the velocity space. There are a number of ways to enrich the space. For example, the unstable linear velocity–linear pressure Stokes element may be stabilized by the addition of a single internal velocity degree of freedom via a bubble. See Figure 2. This is the MINI element of Arnold, Brezzi, and Fortin [3]. A related element is the quadratic velocity–linear pressure Stokes element or Taylor–Hood element. By passing to quadratic velocities, not only is the element stable (on all but some very special mesh topologies), but a higher rate of convergence is achieved. The Taylor–Hood element was conceived independent of any proof of its stability, and verifying stability is much more difficult than for the MINI element or any of the other Stokes elements discussed in this section.

A second (closely related) method of enrichment is to use a finer mesh for velocity than pressure. For example, although the quadrilateral bilinear velocity–constant pressure

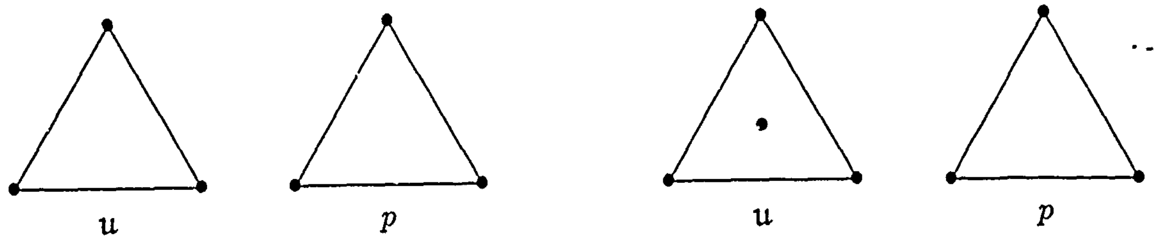


FIG. 2. *An unstable Stokes element (left), stabilized by a bubble degree-of-freedom (right).*

element is unstable (giving rise to the famous checkerboard pressure modes), it can be stabilized by using a composite velocity element which is bilinear on each of four quadrilateral microelements for each quadrilateral pressure element. See Figure 3.

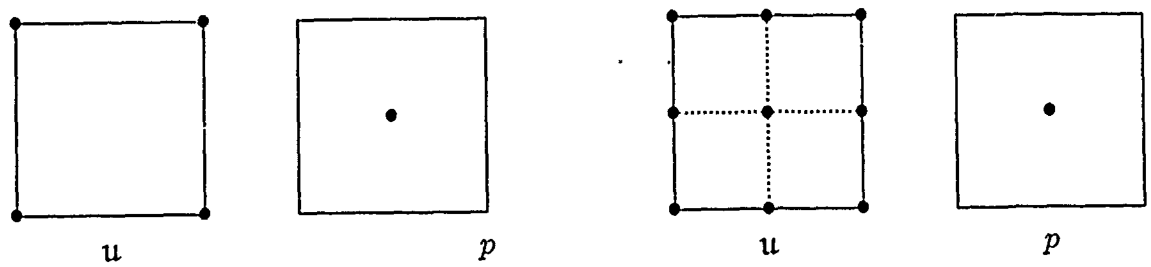


FIG. 3. *An unstable Stokes element (left), stabilized by use of a composite velocity element (right).*

Yet another method to stabilize a Stokes element is to enrich a conforming velocity space to a non-conforming one. This is the idea behind the Crouzeix–Raviart [11] method shown on the right hand side of Figure 4.

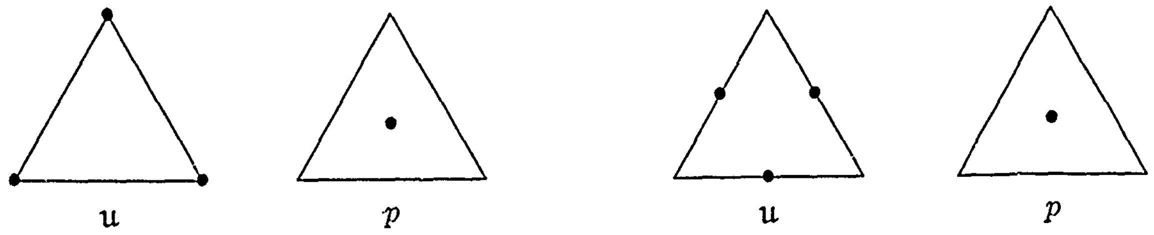


FIG. 4. *Conforming linear velocity–constant pressure (left) is unstable, but non-conforming linear velocity–constant pressure (right) is stable.*

Practically any Stokes element can be rendered stable by sufficient enrichment of the velocity space, the only limitation being the cost associated to the extra velocity degrees of freedom. If the original element afforded similar rates of approximation of velocity and pressure, the additional degrees of freedom for velocity won't increase the rate of convergence, since the approximation will be limited by the pressure. Thus, for instance, the quadratic velocity–constant pressure element pictured in Figure 1b rather disappointingly converges at the rate of the best approximation by *linear* velocity and constant pressure.

For problems in which the a form is not coercive (such as our other three examples), enrichment of the dual variable space need not improve stability. The difficulty is that as S_h gets larger, so does Z_h, and hence the first condition (8) becomes *harder* to satisfy. Thus we are confronted with a trade-off that didn't appear for the Stokes problem. While bubbles, composite elements, and non-conforming elements have all been used in the development of elements for elasticity and other problems, their use is more subtle than for the Stokes problem. For example, Figure 5 shows two stable elements for plane elasticity, the first due to Johnson and Mercier [17], the second to Arnold, Douglas, and Gupta [4], each of which utilizes a composite element for the stress.

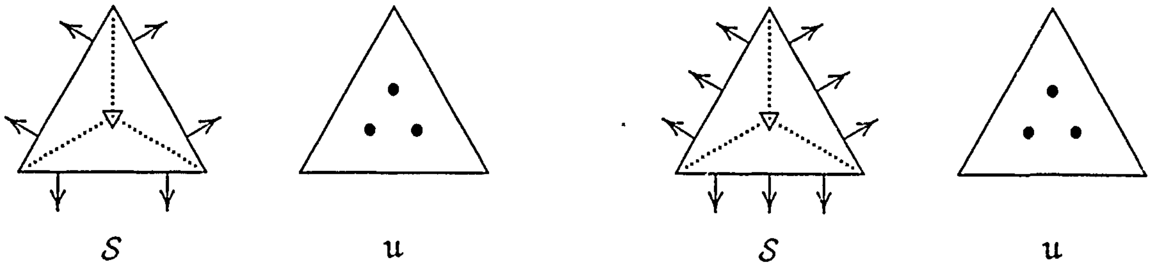


FIG. 5. *The Johnson–Mercier and Arnold–Douglas–Gupta elasticity elements, both of which are stable. The arrows represent the traction (two components). The triangles represent the stress tensor (three independent components). For the former element the stresses are piecewise linears with certain continuity across the dotted lines; for the latter piecewise quadratics are used.*

Many mixed finite elements utilize the fact that the dual variable in three of our example problems is required to have square integrable divergence, but the entire gradient need not exist. That is, the dual variable is sought in $H(\mathrm{div})$ not in the smaller space H^1. For a piecewise polynomial function to belong to $H(\mathrm{div})$ it is not necessary that it be continuous across interelement boundaries (as it would have to be were it to lie in H^1). Only the normal component need be continuous—the tangential component is unconstrained. Thus finite element functions which are discontinous may nevertheless be conforming approximations of $H(\mathrm{div})$. This allows a certain flexibility which can be exploited in the construction of elements. Thus for example the Raviart–Thomas elements and the Brezzi–Douglas–Marini elements, the simplest cases of which are pictured in Figure 6, are of this sort. The latter element can be thought of as a means to stabilize the continuous linear flux–constant temperature element, which is unstable. It also uses linear elements for flux and constant elements for temperature, but the space for fluxes is increased by allowing functions with only the normal component continuous.

The flexibility afforded by using elements which are discontinuous in some components has also been applied to plate bending problems. The variational principle for the Kirchhoff–Love plate can be set up so that continuity is only required on the normal bending moment, $n^t\mathcal{M}n$, while the tangential and twisting moments can jump. The Hellan–Herrmann–Johnson element exploits this flexibility to enable the use of a piecewise constant approximation to the moment tensor. This element, which is diagrammed in

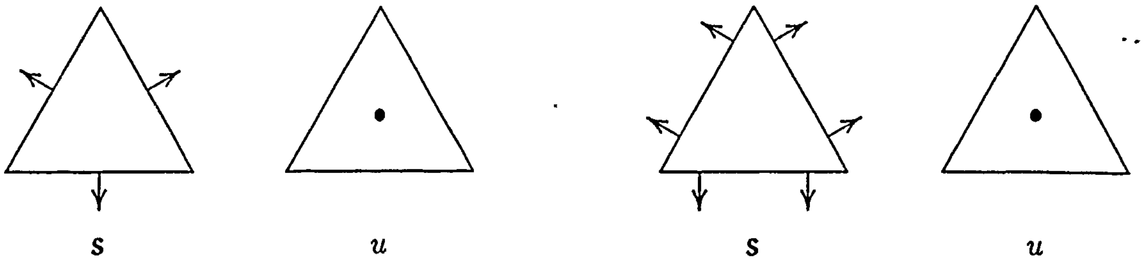


FIG. 6. *The lowest order cases of the Raviart–Thomas and Brezzi–Douglas–Marini elements for scalar second order elliptic problems. The arrows represent the normal component of flux. The flux space for the Raviart–Thomas element consists of vectorfields of the form $(a + bx, c + by)$ on each element. The flux space for the Brezzi-Douglas-Marini space is the full space of linear vectorfields on each element.*

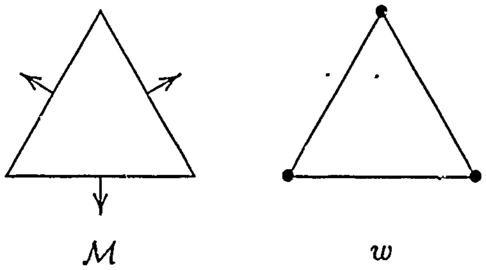


FIG. 7. *The Hellan–Herman–Johnson plate bending element. The moment tensor is approximated by a piecewise constant function. The arrows represent the normal bending moment component on the edge. The transverse displacement is approximated by a continuous piecewise linear function.*

Figure 7, has been shown to be stable in a specially devised, mesh-dependent norm [7].

Most of the examples presented above have been of elements of quite low order. In fact, it is usually easier to obtain stability with higher order elements. Thus, for example, the rather natural Stokes element based on continuous velocity elements of degree d and discontinuous pressure elements of degree $d - 1$ is unstable for $d = 1, 2$, or 3, but can be shown to be stable for $d \geq 4$ [21]. (If discontinuous pressure elements are used, then stable elements are achieved for all degrees $d > 1$, the case $d = 2$ being the Taylor–Hood element.)

The stability constant depends jointly on the bilinear form B (which, for mixed methods, is built of the two forms a and b), and the trial space V_h (which, for mixed methods, is built of the two spaces S_h and W_h). Thus far we have discussed ways to construct and modify the trial space for a given bilinear form in order to obtain stability. It is possible to modify the bilinear form as well. For example, while the linear velocity–linear pressure Stokes element is not stable with the usual bilinear form, if the discrete solution is determined from the system

$$\int_\Omega [C\mathcal{E}(u_h) : \mathcal{E}(v) + \operatorname{div} v \, p_h + \operatorname{div} u_h \, q - h^2 \operatorname{grad} p_h \operatorname{grad} q] = \int_\Omega f \cdot v,$$

the resulting method is stable. This is the simplest example of a Galerkin–Least Squares method.* In this case the change in the bilinear form introduces a consistency error (which however is small enough not to affect the rate of convergence). It is also possible to modify the bilinear form in a consistent way and still have the linear/linear element stable [16]. In the last five years there have been numerous papers presenting extensions and variations of this procedure to obtain simple, stable mixed methods for a variety of problems.

An alteration of the mixed variational formulation for elasticity of an entirely different sort was introduced by Arnold and Falk [5]. They derived a variational principle involving the displacement field and a second-order tensorfield called the pseudostress, from which the true stress can easily be recovered as a linear combiniation of components. Their new variational principle is very similar to the Hellinger–Reissner principle but does not require a symmetry constraint on the tensorfield. This allows one to easily adapt mixed elements for the scalar second order elliptic problem, such as the Raviart–Thomas or Brezzi–Douglas–Marini elements described above.

REFERENCES

[1] D. N. ARNOLD, *Discretization by finite elements of a model parameter dependent problem*, Numer. Math., 37 (1981), pp. 405–421.

[2] D. N. ARNOLD AND F. BREZZI, *Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates*, Math. Modelling and Numer. Anal., 19 (1985), pp. 7–32.

[3] D. N. ARNOLD, F. BREZZI AND M. FORTIN, *A stable finite element for the Stokes equations*, Calcolo, 21 (1984), pp. 337–344.

[4] D. N. ARNOLD, J. DOUGLAS AND C. GUPTA, *A family of higher order mixed finite element methods for plane elasticity*, Numer. Math., 45 (1984), pp. 1–22.

[5] D. N. ARNOLD AND R. S. FALK, *A new mixed formulation for elasticity*, Numer. Math., 53 (1988), pp. 13–30.

[6] I. BABUŠKA AND J. E. OSBORN, *Generalized finite element methods: their performance and their relation to mixed methods*, SIAM J. Numer. Anal., 20 (1983), pp. 510–536.

[7] I. BABUŠKA, J. E. OSBORN AND J. PITKÄRANTA, *Analysis of mixed methods using mesh-dependent norms*, Math. Comp., 35 (1980), pp. 1039–1062.

[8] K. J. BATHE AND F. BREZZI, *On the convergence of a four-node plate bending element based on Mindlin/Reissner plate theory and a mixed interpolation*, in Mathematics of Finite Elements and Applications V, J. R. Whiteman, ed., Academic Press, New York, NY, 1985, pp. 491–503.

[9] J. BOLAND AND R. NICOLAIDES, *Stability of finite elements under divergence constraints*, SIAM J. Numer. Anal., 20 (1983), pp. 722–731.

[10] F. BREZZI, *On the existence, uniqueness, and approximation of saddle point problems arising from Lagrangian multipliers*, RAIRO Anal. Numér., 8–32 (1974), pp. 129–151.

[11] M. CROUZEIX AND P. -A. RAVIART, *Conforming and non conforming finite element methods for solving the stationary Stokes equations*, RAIRO Anal. Numér., R3 (1973), pp. 33–76.

[12] B. DARLOW, R. EWING AND M. WHEELER, *Mixed finite element methods for miscible displacement in porous media*, Sixth SPE Symposium on Reservoir Simulation, SPE 10501, New Orleans, 1982.

*If the bubble degree of freedom is eliminated from the MINI element by static condensation, one arrives, essentially, at the same method.

[13] R. EWING, T. RUSSELL AND M. WHEELER, *Simulation of miscible displacement using mixed methods and a modified method of characteristics*, Seventh SPE Symposium on Reservoir Simulation, SPE 12241, San Francisco, 1983.

[14] M. FORTIN, *Calcul numérique des écoulements des fluides de Bingham et des fluides Newtoniens incompressible par la méthode es éléments finis*, Univ. Paris, Thesis, 1972.

[15] B. X. FRAEIJS DE VEUBEKE, *Displacement and equilibrium models in the finite element method*, in Stress Analysis, O. C. Zienkiewicz and G. Hollister, eds., John Wiley & Sons, New York, NY, 1965.

[16] T. J. R. HUGHES, L. P. FRANCA AND M. BALESTRA, *A new finite element formulation for computational fluid mechanics: V. Circumventing the Babuška–Brezzi condition: a stable Petrov-Galerkin formulation of the Stokes problem accounting for equal order interpolation*, Comput. Methods Appl. Mech. Engrg., 59 (1986), pp. 85–99.

[17] C. JOHNSON AND B. MERCIER, *Some equilbrium finite element methods for two-dimensional elasticity problems*, Numer. Math., 30 (1978), pp. 103–116.

[18] C. JOHNSON AND J. PITKÄRANTA, *Analysis of some mixed finite element methods related to reduced integration*, Math. Comp., 38 (1982), pp. 375–400.

[19] D. S. MALKUS AND T. J. R. HUGHES, *Mixed finite element methods—reduced and selective integration techniques: a unification of concepts*, Comput. Methods Appl. Mech. Engrg., 15 (1978), pp. 63–81.

[20] L. D. MARINI AND A. SAVINI, *Accurate computation of electric field in reverse-biased semiconductor devices: a mixed finite element approach*, Compel, 3 (1984), pp. 123–135.

[21] L. R. SCOTT AND M. VOGELIUS, *Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 111–143.

Adaptive finite element methods for diffusion and convection problems

Claes Johnson
Mathematics Department
Chalmers University of Technology
412 96 Göteborg, Sweden

Abstract. We give a survey of recent results obtained together with K. Eriksson on adaptive h-methods for the basic linear partial differential equations of elliptic, parabolic and hyperbolic type. Our adaptive algorithms are based on a posteriori error estimates leading to reliable methods, and comparison with sharp a priori error estimates is made to prove efficiency of the procedures.

Introduction.

In this note we give a survey of some recent results on adaptive finite element methods obtained in collaboration with Kenneth Eriksson, (see [E1-3], [EJ1-5]). As model problems we shall consider the heat equation including the corresponding stationary Poisson equation representing diffusion-dominated problems, and also linear convection-dominated convection-diffusion problems. Together, these problems cover the basic linear partial differential equations of parabolic, elliptic and (first order) hyperbolic type. In each of these cases our goal is to solve the following problem (A): Given a norm $\|\cdot\|$, a tolerance $TOL > 0$, and a piecewise polynomial finite element discretization of a certain type (e.g. piecewise polynomials of a certain given degree), design an algorithm for constructing a mesh T with (nearly) minimal number of degrees of freedom, such that

$$\|u-U\| \leq TOL, \tag{0.1}$$

where u is the exact solution and U is the finite element solution on the mesh T. Clearly, our problem (A) has two ingredients: First, we want the adaptive algorithm to be reliable in the sense that the error control (0.1) is guaranteed. Secondly, we want the algorithm to be efficient in the sense that the constructed mesh is nowhere overly refined. Note that for definiteness our criterion for efficiency is a minimal number of degrees of freedom. Of course, in practice depending on the particular implementation, mesh generator, solution techniques et cet., we may accept a certain over-refinement.

Adaptive codes are now entering into applications and adaptivity may be expected to become a standard feature of finite element software in the future. Quantitative error control is of obvious interest in applications and efficient techniques for adaptive local refinement or orientation of the mesh opens fascinating possibilities of computing accurate solutions to complex problems involving different scales, such as problems in fluid mechanics with boundary layers and shocks, crack problems in solid mechanics, semiconductor problems, reaction-diffusion problems et. cet.

Our adaptive algorithms are based on a posteriori error estimates of the form

$$(0.2) \qquad \|u-U\| \leq \mathcal{E}(U, h, \text{data}),$$

where as indicated the error bound $\mathcal{E}$ depends on the computed solution U, on the mesh size h of the corresponding mesh T and the data of the problem. Here h is a function giving the local mesh size in space and time. Starting from (0.2) we have the following adaptive method for error control in the $\|\cdot\|$-norm to the tolerance TOL: Find a mesh T, with mesh function h and corresponding approximate solution U, with minimal number of degrees of freedom such that $\mathcal{E}(U, h, \text{data}) \leq$ TOL. Since U depends on h this is a (complex) non-linear minimization problem. To solve this minimization problem approximately we design an adaptive algorithm usually of the following form: Given a first coarse mesh T_0, construct successively meshes T_j, j = 1,...,J, with corresponding mesh functions h_j and approximate solutions U_j, with minimal number of degrees of freedom such that:

$$(0.3) \qquad \mathcal{E}(U_{j-1}, h_j, \text{data}) \leq \theta\,\text{TOL},$$

until $\mathcal{E}(U_J,h_J, \text{data}) \leq$ TOL, which is the stopping criterion. Here θ is factor ($\theta \sim 1$) influencing the total number of steps J required to reach the stopping criterion. Since

U_{j-1} is given in (0.3), the minimization problem in h_j is easy to solve approximately by seeking to equidistribute the element contributions to the global quantity $\mathcal{E}$. Note that we consider here adaptive forms of the so called h-method, where the quantity determined adaptively is the local element size. More generally, it is of interest to develop methods where the mesh orientation and stretching, and the degree of the piecewice polynomials are also determined adaptively.

Since our adaptive algorithms are based on a posteriori error estimates, it follows that the algorithms are reliable in the above sense; if the stopping criterion $\mathcal{E}(U,h, \text{data}) \leq \text{TOL}$ is satisfied, then by (0.2) we will have $\|u-U\| \leq \text{TOL}$ and the error will be within the given tolerance. Concerning the efficiency of the adaptive algorithms more or less precise results may be stated. Ideally, one would like to prove that the final mesh generated by the algorithm is close to the optimal mesh, which we take to be the mesh with fewest degrees of freedom such that the approximate solution is within the given tolerance. In certain cases it is possible to actually prove such a precise result (up to constants of moderate size), while in other cases we obtain weaker results. In general, to prove efficiency we rely on sharp a priori error estimates, where the error $\|u-U\|$ is estimated by a quantity $E(u,h)$ depending on the exact solution and the mesh size h. In the elliptic and parabolic case we prove that the a posteriori quantity $\mathcal{E}(U,h,\text{data})$ may be estimated by a constant times the a priori quantity $E(u,h)$, which proves efficiency in a weak sense, and may be sharpened through various localization results to prove efficiency in a strict sense for certain problems. Note that by proving that the a posteriori bounds may be estimated by a constant times the a priori bounds, it follows in particular that by decreasing the mesh size it is possible to realize the stopping criterion under some assumption on the nature of the exact solution, which is not evident from the start.

Summing up so far, our adaptive algorithms are based on (sharp) a posteriori error estimates leading to reliable methods, and comparison with sharp a priori error estimates is used to prove efficiency in a more or less precise way. The error estimates are based on a representation of the error in terms of the solution of a certain dual problem. This error representation is fundamental in our approach to adaptivity since it gives information on the structure of the global error as composed of contributions from individual elements, which gives the basis for the design of the adaptive algorithm. .. Basically, the error estimates are obtained by using the orthogonality properties of the Galerkin method and standard finite element interpolation estimates, together with appropriate stability estimates for the dual problem. In the case of the a posteriori error estimates the dual problem is a continuous problem, while for the a priori estimates the dual problem is discrete. In this framework there is a close analogy between the a priori and a posteriori error estimates which appears to be fundamental. In both cases the stability of the dual problem is the critical issue. In general, the stability of the continuous dual problem connected with the a posteriori estimates is easier to tackle by analytical tools than that of the discrete dual problem related to the a priori estimates, and thus in many cases the a posteriori estimate is easier to prove than the a priori estimate, contrary to a common opionion that a posteriori estimates are more difficult to obtain. For more general problems (e.g. nonlinear problems or problems with variable coefficients) the stability of the continuous dual problem cannot be accurately estimated by analytical means (in particular, one has to determine approximiately the size of certain constants involved) and in these cases one has to build in a computational estimate of the stability of the dual problem as a part of the adaptive process. For certain problems and norms, numerical experiments show that this is feasible, while for more general problems more work is required to obtain reliable and accurate computational estimates of the stability of the dual problem. We note that this is a fundamental problem which has to be faced, analytically or computationally, since the stability of the dual problem reflects the error propagation properties of the given equation.

For parabolic problems we use the Discontinuous Galerkin method based on a space-time finite element discretization with basis functions continuous in space and discontinuous in time. The time step and the space discretization may vary from one time level to the next and it is also possible to use more general space-time meshes with the time steps being variable also in space. For hyperbolic type problems we use the Streamline Diffusion method (SD-method for short below) again with space-time finite elements in the time-dependent case.

We now briefly comment on the difference in our approach to adaptivity as compared to the pioneering work by Babuska, see e.g. [B], and the related work by Bank [Ba1] and Ewing [Ew]. First, in the work by Babuska et al the emphasis is on elliptic problems with error control in the energy norm, while we consider also other norms and also parabolic and hyperbolic problems. Secondly, Babuska et al seek to construct a posteriori error bounds which are very precise in the sense that the quotient between the estimated and the actual global error (the effectivity index) tends to one as the mesh size tends to zero. For this purpose elaborate a posteriori estimates based on solving local problems are used. However, in our approach we set the goal lower in this respect, and we use simpler possibly less precise estimates and accept (depending on the difficulty of the problem, the chosen norm and the tolerance) effectivity indices in the range, say, from 1 to 3. In contrast we are able to attack more general problems and we are not restricted to only energy norms. A further difference is that we seek to obtain adaptive algorithms which we can prove to be efficient in a more or less precise way. Let us note that we should distinguish between the concept of effectivity index and the efficiency of the adaptive algorithm. Even if the effectivity index is close to one, which says that we are able to estimate the global error on a given mesh very accurately, it is not clear that the underlying mesh is close to the optimal mesh related to the corresponding tolerance level; the given mesh may be locally overly refined and there is no way we can detect this by only looking at the effectivity index.

To sum up, in our approach we do not seek to achieve effectivity indices necessarily very close to one, but we seek adaptive algorithms for a general class of problems with error control in a variety of norms and we seek to prove that the algorithms are efficient in the sense that almost optimal not overly refined meshes are generated. Note that for parabolic problems our adaptive methods seem to be the first to give reliable and efficient error control in $L_\infty(L_2)$ i.e. the maximum norm in time and L_2

in space. For hyperbolic problems our results appear to give the first adaptive methods based on a posteriori error estimates.

After our paper [EJ3] was completed we discovered that our a posteriori error estimates in the energy norm (H^1-norm) for the Poisson equation are analogous to those presented in Abdalass [A] and Verfürth [V] for the Stokes problem. We also learned that similar a posteriori error estimates for the Poisson equation were considered already 1979 by Bank [Ba2]. These estimates are based on estimating the H^{-1}-norm of the residual of the finite element solution in terms of a weighted L_2-norm of the residual over the element interiors and the jumps of the normal derivatives of the finite element solution across interelement boundaries (using the orthogonality relation built in the Galerkin method). In this approach (which is very simple and natural) the residual of the finite element solution is separately estimated in the interior of the elements and on element boundaries, which leads to effectivity indices not necessarily very close to one. The development in the early and mid eighties with mathematical emphasis, however, took a different route concentrating on methods with effectivity index close to one. Our interest in a posteriori estimates of the indicated form is motivated by their simplicity and the possibility of handling problems of different nature and different norms. For pioneering work on adaptive methods with emphasis on engineering aspects, see also [LMZ], [ODSD].

An outline of the remainder of this notes is as follows: In Section 1 we present the discretization methods for our model problems of elliptic, parabolic and hyperbolic type. In Section 2 we present the a priori and a posteriori error estimates and in Section 3 we state the corresponding adaptive algorithms and in Section 4 we discuss their reliability and efficiency. In Section 5 we indicate the structure of the proofs of the a posteriori and a priori error estimates and finally in Section 6 we present the results of some numerical experiments.

1. The discretization methods.

For simplicity we shall restrict our considerations to some standard model problems of elliptic, parabolic and hyperbolic type, namely, to find u such that

$$(1.1)\qquad \begin{cases} -\Delta u(x) = f & \text{in } \Omega, \\ u(x) = 0 & \text{on } \Gamma, \end{cases}$$

$$(1.2)\qquad \begin{cases} u_t - \Delta u = f & \text{in } \Omega \times \mathbb{R}^+, \\ u = 0 & \text{on } \Gamma \times \mathbb{R}^+, \\ u(\cdot, 0) = u_0 & \text{in } \Omega, \end{cases}$$

$$(1.3)\qquad \begin{cases} \beta \cdot \nabla u + \alpha u - \operatorname{div}(\epsilon \nabla u) = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma, \end{cases}$$

$$(1.4)\qquad \begin{cases} u_t + \beta \cdot \nabla u + \alpha u - \operatorname{div}(\epsilon \nabla u) = f & \text{in } \Omega \times \mathbb{R}^+, \\ u = g \quad \text{on } \Gamma \times \mathbb{R}^+, \\ u(\cdot, 0) = u_0 \quad \text{in } , \end{cases}$$

respectively. Here Ω is a bounded polygonal domain in $\mathbb{R}^2$ with boundary Γ, $\mathbb{R}^+ = (0,\infty)$, Δ is the usual Laplacian , $u_t = \frac{\partial u}{\partial t}$, $\beta = (\beta_1, \beta_2)$ is a given velocity field, α is a given absorption coefficient, $\epsilon \geq 0$ a given (small) diffusion coefficient, all coefficients possibly depending on x and t, and the functions f, g and u_0 are given data. We note that when $\epsilon = 0$, then the boundary condition $u = g$ in (1.3) is imposed only on the inflow part of the boundary $\Gamma_- = \{x \in \Gamma : n(x) \cdot \beta(x) < 0\}$ where $n(x)$ is the outward unit normal to Γ at $x \in \Gamma$, and similarly for (1.4).

For the discretization of these problems with respect to the space variable $x = (x_1, x_2)$, let Σ be the class of all finite element discretizations (h, T, S) defined as follows: (i) h is a positive function in $C^1(\overline{\Omega})$ such that

$$(1.5a)\qquad |\nabla h(x)| \leq \mu, \quad \forall x \in \overline{\Omega},$$

(ii) $T = \{K\}$ is a triangulation subdomains of Ω into triangles K of diameter h_K such that

(1.5b) $\quad c_1 h_K^2 \leq \int_K dx, \ \forall K \in T,$

and associated with the function h through

(1.5c) $\quad c_2 h_K \leq h(x) \leq h_K, \ \forall x \in K, \ \forall K \in T,$

where c_1 and c_2 are given positive constants which together with μ and Ω characterize Σ, and (iii) S is the set of all continuous functions on $\overline{\Omega}$ which are linear in x on each $K \in T$ and vanish on $\partial\Omega$.

As indicated, in the adaptive process we need to construct for a given mesh function h satisfying (1.5a) a corresponding mesh T satisfying (1.5b,c). In our implementations we have for this purpose used two mesh generators: one based on successive subdivision of one triangle into four similar triangles by joining the midpoints of the sides of the given triangle and introducing "transition triangles" divided into two subtriangles connceting zones with different mesh size, and another "front generator" which constructs a mesh with given local mesh size by adding elements at a "front" coinciding initially with the boundary and sweeping the region, see [H].

The stationary elliptic problem (1.1) may now be approximated in the usual way: Let $(h, T_h, S_h) \in \Sigma$ and find $U \in S_h$ such that

(1.6) $\quad (\nabla U, \nabla v) = (f,v), \quad \forall v \in S_h,$

where $(\cdot,\cdot)$ denotes the usual inner product in $[L_2(\Omega)]^d$, d = 1,2. By introducing the L_2 projection operator $P_h: L_2(\Omega) \to S_h$ defined by $(P_h w,v) = (w,v), \ \forall v \in S_h$, and the discrete Laplacian $\Delta_h: H^1(\Omega) \to S_h$ defined by $(\Delta_h w,v) = -(\nabla w, \nabla v), \ \forall v \in S_h$, we may write (1.6) equivalently as $-\Delta_h u_h = P_h f$, which has a more obvious resemblance with (1.1).

Let us now turn to the time dependent parabolic problem (1.2). For a full discretization of this problem with the Discontinuous Galerkin method we consider partitions $0 = t_0 < t_1 <\ldots< t_n <\ldots$, of $\mathbb{R}^+$ into subintervals $I_n = (t_{n-1}, t_n)$ of length $k_n = t_n - t_{n-1}$, and associate with each such time interval a space discretization $(h_n,$

$T_n, S_n) \in \Sigma$. For q a nonnegative integer we define $V_{qn} = \{v: v = \sum_{j=0}^{q} t^j \varphi_j, \varphi_j \in S_n\}$, and discretize (1.2) as follows: Find U such that for $n = 1,2,\ldots$, $U|_{\Omega\times I_n} \in V_{qn}$ and

$$\int_{I_n} \{(U_t,v) + (\nabla U,\nabla v)\}dt + ([U]_{n-1}, v^+_{n-1}) = \int_{I_n} (f,v)dt, \quad \forall v \in V_{qn}, \tag{1.7}$$

where $[w]_n = w_n^+ - w_n^-$, $w_n^{+(-)} = \lim_{s\to 0^{+(-)}} w(t_n+s)$, and $U_0^- = u_0$. With $f = 0$, (1.7) is equivalent to the sub-diagonal $(q+1, q)$-Padé scheme of order of accuracy $2q + 1$, see [EJT].

Remark: Note that in the discretization (1.7) the space and time steps may vary in time and that the space discretization may be variable also in space, whereas the time steps k_n are kept constant in space. Clearly, optimal mesh design requires the time steps to be variable also in space. Now, it is easy to extend the method (1.7) to admit time steps which are variable in space simply by defining

$$V_{qn} = \{v: v(x,t) = \sum_i v_i(t)\chi_i(x)\},$$

where $\{\chi_i\}$ is a basis for S_n and the coefficients v_i now are piecewise polynomial of degree q in t, without continuity requirements, on partitions of I_n which may vary with i. The discrete functions may now be discontinuous also inside the "slab" $\Omega \times I_n$. The Discontinuous Galerkin method again takes the form (1.7) with the difference that the term $([U]_{n-1}, v^+_{n-1})$ is replaced by a sum over all jumps of U in $\Omega\times[t_{n-1}, t_n)$ and further the discontinuities of U_t are discarded in the integral involving U_t. Adaptive methods for the Discontinuous Galerkin method in this generality are considered in [EJ4]. ■

Finally, we consider the convection diffusion problems (1.3) and (1.4). For the discretization of these problems we shall as indicated use the SD-method which is a variant of a standard Galerkin finite element method obtained by two basic modifications: a "streamline" modification of the test functions (in the stationary case) from v to $v + \delta(\beta\cdot\nabla v + \alpha v)$ where $\delta \sim h$, and a second modification obtained by adding an artificial viscosity term with viscosity coefficient proportional to h^{α} (with $3/2 < \alpha < 2$) and the residual of the finite element solution. The streamline SD-method is the first general finite element method for (first order) hyperbolic equations which combines good stability with higher order accuracy. Convergence results are available for linear scalar convection-diffusion problems, for the incompressible Euler and Navier-Stokes equations, for scalar conservation laws in several dimensions, and also (entropy) consistency results for e.g. the compressible Euler and Navier-Stokes equations (see [JNP], [JSW], [JSH], [S]). With $q = 1$ the SD-method for (1.3) may be formulated as follows in the case $g = 0$ and $\epsilon > 0$: Find $U \in S_h$ such that

$$(1.10) \qquad (\beta\cdot\nabla U + \alpha U, v + \delta(\beta\cdot\nabla v + \alpha v)) + (\hat{\epsilon}\nabla U, \nabla v) = (f, v + \delta(\beta\cdot\nabla v + \alpha v)) \quad \forall v \in S_h,$$

where

$$\delta = C_1 \max(h - \frac{\epsilon}{|\beta|}, 0)/|\beta|,$$
$$\hat{\epsilon} = \hat{\epsilon}(U) = \max(\epsilon, C_2 h^{\alpha}|\beta\cdot\nabla U + \alpha U - f|),$$

where the C_i and α are positive constants with $\frac{3}{2} < \alpha < 2$. In the computations we normally choose α close to 2.

For the time dependent problem (1.4) the SD-method reads as follows using the notation of (1.7) again with $q = 1$ and assuming that $g = 0$ and $\epsilon > 0$: Find U such that for $n = 1,2,\ldots,$ $U|_{\Omega\times I_n} \in V_{1n}$ and

$$(1.11) \qquad \int_{I_n} \{(U_t + \beta\cdot\nabla U + \alpha U, v + \delta(v_t + \beta\cdot\nabla v + \alpha v))\}dt$$

$$+ \int_{I_n} (\hat{\epsilon}\hat{\nabla}U, \hat{\nabla}v)dt + ([U_{n-1}], v_{n-1}^{+})$$

$$= \int_{I_n} (f,v + \delta(v_t + \beta\cdot\nabla v + \alpha v))dt, \quad \forall v \in V_{1n},$$

where $U_0^- = u_0$,

$$\delta = C_1 \ \max (h - \frac{\epsilon}{|(1,\beta)|},0)/|(1,\beta)|,$$

$$\hat{\epsilon} = \hat{\epsilon}(U) = \max(\epsilon, C_2 h^{\alpha}(|U_t+\beta\cdot\nabla U + \alpha U - f| + |\frac{[U_{n-1}]}{k_n}|)) \ \text{in} \ \Omega \times I_n,$$

$$\hat{\nabla}v = (\frac{\partial v}{\partial t}, \frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2}),$$

with the C_i and α as above.

2. A priori and a posteriori error estimates.

In this section we state a priori and a posteriori error estimates for the discretization methods (1.6)-(1.11). By $\|\cdot\|$ we denote the $L_2(\Omega)$-norm and $D^s u = \left[\sum_{|\alpha|=s} |D^{\alpha}u|^2\right]^{1/2}$. For the stationary elliptic problem (1.1) we have the following a priori estimate:

Theorem 2.1. Let $f \in L_2(\Omega)$ and let u and U be the solutions of (1.1) and (1.6), respectively. Then for $m = 1$ and 2 there exists a constant C depending only on the constants c_1 and c_2 in (1.5), such that

$$\|D^{2-m}(u-U)\| \le C\|h^m D^2 u\|, \tag{2.1}$$

where for $m = 2$ we assume that μ is sufficiently small and Ω is convex.

Remark 2.1. Note the way in which the local mesh size $h(x)$ enters in these error estimates showing that large second derivatives of u may be compensated for by a (locally) small mesh size so as to control the quantity $\|h^m D^2 u\|$, $m=1,2$, bounding the error. This indicates the possibility of adaptively choosing the mesh size to control the error if D^2u may be computationally estimated, an idea which was explored in [EJ1]. In this note, however, we will follow a related but different adaptive strategy directly based on a posteriori error estimates. ∎

Remark 2.2: Note further that the estimates (2.1a,b) are optimal in the sense that there exists a constant c such that "for most u" (e.g. if $D^{\alpha}u$, $|\alpha| = 2$, is roughly constant on each element),

$$\inf_{v \in S_h} \|D^{2-m}(u-v)\| \geq c\,\|h^m D^2 u\|, \qquad m = 1,2,$$

which indicates that error control based on (2.1) should be efficient. ∎

The error estimate (2.1) with $m = 1$ is classical, whereas with $m = 2$ the estimate in the present generality can be found in [E2]. For quasi-uniform partitions (corresponding to taking h constant) the case $m = 2$ is well known. Let us further remark that (2.1) may also be derived for a (convex or nonconvex) domain Ω with smooth boundary, in the case $m = 2$ with the constant C depending on Ω.

To state the a posteriori estimate for the stationary elliptic problem (1.1) underlying the adaptive algorithm we need some notation. With each side $\tau \in \partial K \cap \Omega$ of a triangle $K \in T_h$ we associate a vector n_τ of length one normal to τ and define for $v \in S_h$

$$\left[\frac{\partial v}{\partial n_\tau}\right] = \lim_{s \to 0^+} (\nabla v(x+sn_\tau) - \nabla v(x-sn_\tau)) \cdot n_\tau, \quad x \in \tau,$$

that is, $[\frac{\partial v}{\partial n_\tau}]$ is the jump across τ in the normal component of ∇v. We define for $v \in S_h$ the piecewise constant quantity $D_h^2 v$ by

$$D_h^2 v = \max_{\tau \in \partial K \cap \Omega} |[\frac{\partial v}{\partial n_\tau}]|/h_K \quad \text{on } K \in T_h,$$

where as indicated only the sides τ in the interior of Ω occur, which may be viewed as a discrete counterpart of $|D^2v|$.

We may then state the following a posteriori error estimates for the stationary elliptic problem (1.1) given in [EJ3]:

Theorem 2.2. There are constants α_m and β_m only depending on the constants c_1 and c_2 such that if $f \in L_2(\Omega)$ and u and U are the solutions of (1.1) and (1.4), respectively, then for $m = 1,2$,

$$\|D^{2-m}(u-U)\| \leq \alpha_m\|h^m f\| + \beta_m\|h^m D_h^2 U\|, \tag{2.2}$$

where in the case $m = 2$, we assume that μ is sufficiently small and Ω is convex.

Remark 2.4. Note that without further analysis the amount of information in the a posteriori estimate (2.2) is not obvious. Clearly, if we compute U using (1.6), then we may bound the error using (2.2) by evaluating the right hand sides of these estimates. If these quantities turn out to be sufficiently small, then we may be satisfied and quit. However, without further analysis it is conceivable that the right hand sides of (2.2) would always be large and then these estimates would be useless. In fact, a posteriori error estimates of the form (2.2) may be derived also for unstable methods and in such cases the right hand side quantities could be large regardless of the mesh size. In our case we shall prove that in fact (2.2) is sharp, and thus may be useful in practice, by comparson with the optimal a priori estimates (2.1). ■

Remark 2.5. If $f \in H^2(\Omega)$, then the f-terms in (2.2) or (2.3) may be replaced by $\alpha_3\|h^{4-m}D^2 f\|$. ■

Let us next state optimal a priori estimates for the parabolic problem (1.2) given in [EJ3]. For simplicity we assume that Ω is convex. The estimates may be extended to general domains with smooth boundary with the constants C depending on Ω.

Theorem 2.3. Let u be the solution of (1.2) and U that of (1.7), suppose μ is small enough and assume that for each n one of the following two assumptions hold:

$$S_n \subset S_{n-1}, \quad (2.4a)$$

$$\bar{h}_n^2 \le \gamma k_n, \quad (2.4b)$$

where $\bar{h}_n = \max_{x \in \bar{\Omega}} h_n(x)$ and γ is sufficiently small and that for all n, $k_n \le C k_{n+1}$. Then there exist constants C only depending on c_1 and c_2 (if Ω is convex) such that for $q = 0,1,$ and $N = 1,2,\ldots,$

$$\|u-U\|_{I_N} \le CL_N \max_{1 \le n \le N} E_{qn}(u), \quad (2.5a)$$

and for $q = 1$, $N = 1,2,\ldots.,$

$$\|u(t_N) - U_N^-\| \le CL_N \max_{1 \le n \le N} E_{2n}(u), \quad (2.5b)$$

where $L_N = \frac{1}{4}(\ell n \frac{t_N}{k_N} + 1)^{\frac{1}{2}}$,

$$E_{qn}(u) = \min_{j \le q+1} k_n^j \|u_t^{(j)}\|_{I_n} + \|h_n^2 D^2 u\|_{I_n}, \quad q = 0,1,2,$$

with $u_t^{(1)} = u_t$, $u_t^{(2)} = u_{tt}$, $u_t^{(3)} = \Delta u_{tt}$ and $\|w\|_{I_n} = \max_{t \in I_n} \|w(t)\|$.

Remark 2.6. Note that (2.5) states that the Discontinuous Galerkin method (1.5) is of order q+1 globally in time and of order 2q+1 at the discrete time levels t_n for q = 0,1. Further, the estimates (2.5) are optimal in the sense that for some positive constant c

$$(2.6) \qquad \inf_{v\in V_{qn}} \|u-v\|_{I_n} \geq c\, E_{qn}(u), \qquad q = 0,1,2,$$

if here, in the definition of $E_{qn}(u)$, we put $u_t^{(3)} = u_{ttt}$ and restrict the variation of $u_t^{(3)}$ and $D^\alpha u$ for $|\alpha| = 2$ as in Remark 2.2. Note that for the "super approximation" result (2.5b) it is relevant to compare with approximation in V_{2n}. ■

Remark 2.7. With quasi-uniform space-meshes with $h_n(x) \sim \bar{h}_n$ we expect to have $\bar{h}_n^2 \sim k_n$ for $q = 0$ and $\bar{h}_n^2 << k_n$ if $q = 1$, since the Discontinuous Galerkin method is of second order in space and of order $2q+1$ in time, $q = 0,1$. Thus, in particular for $q = 1$ the condition (2.4b) does not appear to be restrictive and in fact allows a considerable variation of $h_n(x)$. In certain extreme situaitions, however, e.g. with initial data u_0 highly concentrated in space, (2.4b) may impose a restriction on the mesh. It is possible that (2.4b) may be weakened to a condition of the form $\bar{h}_n^2 \leq \gamma K_n$, where $K_n = t_{n^*} - t_{n-1}$, and $S_m = S_n$ for $m = n,n+1,\ldots, n^*$. ■

We now state a posteriori estimates for the parabolic problem (1.2) (see [EJ3]. Again, we assume that Ω is convex, but generalizations to smooth non-convex domains are possible (cf. Remark 2.10).

Theorem 2.4. Let u be the solution of (1.2) and U that of (1.5), suppose Ω is convex and μ sufficiently small. Then for $N \geq 1$, we have for $q = 0$

$$(2.7a) \qquad \|u(t_N) - U_N^-\| \leq \max_{1\leq n\leq N} \mathcal{E}_{on}(U),$$

and for $q = 1$

$$(2.7b) \qquad \|u(t_N) - U_N\| \leq \max_{1\leq n\leq N} \mathcal{E}_{2n}(U),$$

where

$$\mathcal{E}_{on}(U) = C_1\|h_n^2 f\|_{I_n} + C_2 \int_{I_n} \|f\|dt + C_3\|h_n^2 D_n^2 U_n^-\| + C_4\|[U_{n-1}]\| + C_5\|h_n^2[U_{n-1}]/k_n\|^*,$$

$$\mathcal{E}_{2n}(U) = C_6\|h_n^2 f\|_{I_n} + C_7 k_n^2 \int_{I_n} \|f_{tt}\|dt + C_8\|h_n^2 D_n^2 U\|_{I_n}$$

$$+ \min(C_9\|[U_{n-1}]\|, C_{10}k_n\|\Delta_n P_n[U_{n-1}]\|) + C_{11}\|h_n^2[U_{n-1}]/k_n\|^*,$$

where $D_n^2 = D_{h_n}^2$, and a star indicates that the corresponding term is present only if $S_n \not\supseteq S_{n-1}$. Further the C_i are constants given by

$$C_1 = \alpha_2 L, \quad C_2 = L+2 \quad C_3 = \beta_2(L+2), \quad C_4 = L+1,$$
$$C_5 = \alpha_2(L+\exp(-1)), \quad C_6 = \alpha_2(L+2), \quad C_7 = \gamma_3(\gamma_1 L + \gamma_0 + 1)$$
$$C_8 = 2\beta_2\,(L+1), \quad C_9 = C_4, \quad C_{10} = \gamma_2 L + \gamma_1, \quad C_{11} = C_5,$$
$$L = \max_N L_N,$$

where α_2 and β_2 are certain constants depending on c_1 and c_2 related to approximation by functions in S_h, and the γ_i are absolute constants related to one-dimensional approximation by linear functions (see Sectio ` below).

<u>Remark 2.8.</u> The term $C_1\|h_n^2 f\|_{I_n}$ in $\mathcal{E}_{on}$ and $\mathcal{E}_{2n}$ may be replaced by $\bar{C}_1\|h_n^4 D^2 f\|_{I_n}$, the term $C_2 \int_{I_n} \|f\|dt$ in $\mathcal{E}_{on}$ by $\bar{C}_2 k_n \int_{I_n} \|f_t\|dt$ and $C_7 k_n^2 \int_{I_n}\|f_{tt}\|dt$ in $\mathcal{E}_{2n}$ by $\bar{C}_7 k_n^3 \int_{I_n} \|\Delta f_{tt}\|dt$ with modified constants $\bar{C}_i$. ■

<u>Remark 2.9.</u> The comments of Remark 2.4 are also relevant for the <u>a posteriori</u> estimate (2.7). By comparison with the optimal <u>a priori</u> estimate (2.5) we can prove that (2.7) is

sharp and thus may be used as a basis for an efficient adaptive algorithm. ■

Remark 2.10. In the general case with the boundary of Ω smooth, (some of) the constants C_i should be replaced by constants $\hat{C}_i = C_s C_i$, where C_s is a stability constant depending on Ω defined by

$$C_s = \sup_{\substack{v \in H_0^1(\Omega) \cap H^2(\Omega) \\ v \neq 0}} \frac{\|D^2 v\|}{\|\Delta v\|} .$$

The approximation constants α_2 and β_2 (depending on c_1 and c_2) and the absolute constants γ_i entering in the C_i, may be estimated once and for all (values of these constants used in our numerical computations are given in Section 9 below), while the stability constant C_s in general depends on Ω. It is possible that a relevant value of C_s may be found by computing the quotient $\|D_h^2 v\|/\|\Delta_h v\|$ for some properly chosen $v \in S_h$. The a posteriori estimates may be generalized also to problems with variable coefficients or nonlinear problems (see [EJ4]). In this case the C_i should be replaced by $\hat{C}_i = C_s(u)C_i$, where $C_s(u)$ is a "stability constant" depending on Ω and the coefficients, and also "mildly" on u. It is likely that such constants may be estimated through the adaptive procedure, cf [E1], [EJ2,4]. ■

Remark 2.11. One can prove direct analogues of Theorem 2.1-4 replacing the $L_2(\Omega)$-norm by the $L_p(\Omega)$-norm, $1 \leq p \leq \infty$, (see [E1]). ■

Finally, we whall state some a priori and a posteriori error estimates for the SD-method for the convection-diffusion problems (1.3) and (1.4). We start with an a priori error estimate from [JNP] for the stationary problem (1.3), for simplicity with α bounded below by a positive constant. Further, for notational simplicity we consider the convection dominated case with $\epsilon < h$. The estimate can easily be extended to a general ϵ to give estimates analogous to (2.2) in the case $\epsilon = 1$.

Theorem 2.5. Suppose there are positive constants κ_i such that $\kappa_0 < \alpha(x) < \kappa_1$, $x \in \bar{\Omega}$, and suppose the velocity β is smooth. If the exact solution u of (1.3) belongs to $W^{1,\infty}(\Omega)$, then there exists a constant C such that if $\epsilon < h$, then

$$\|\delta^{\frac{1}{2}}(\beta\cdot\nabla(u-U))\| + \|\hat{\epsilon}^{\frac{1}{2}}\nabla(u-U)\| + \|u-U\| \leq C\|h^{3/2}D^2u\| .$$

We now state an a posteriori error estimate for the SD-method (1.10) for the stationary problem (1.3) from [EJ4]. For simplicity we shall compare the computed solution U with the solution $\hat{u}$ of a perturbed continuous convection-diffusion problem obtained by replacing ϵ by $\hat{\epsilon}(U)$ in (1.3). It is also possible in model cases to estimate the perturbation error $\|u-\hat{u}\|$ in terms of $\hat{\epsilon}(U) - \epsilon$, U and f, see [EJ5]. In general, we expect $\|u-\hat{u}\|$ to be dominated by $C\|\hat{u}-U\|$, so that control of $\|\hat{u}-U\|$ suffices. In the adaptive algorithm for (1.10) to be presented below, we also have the option of including the requirement $\hat{\epsilon} = \epsilon$, corresponding to resolution of all details of the exact solution, in which case on the final mesh $\hat{u} = u$, see Section 3 below. For simplicity we assume that the coefficients α and β are constant.

Theorem 2.6. There is a constant C such that

$$\|\hat{u} - U\| \leq C\|\min(1, \hat{\epsilon}^{-1}h^2)R(U)\| + \max_{\Gamma_-} \hat{\epsilon}^{\frac{1}{2}} , \tag{2.8}$$

where

$$R(U) = |\beta\cdot\nabla U + \alpha U - f| + |\mathrm{div}_h(\hat{\epsilon}\nabla U)|,$$

$$|\mathrm{div}_h(\hat{\epsilon}\nabla U)| = \max_{\tau\in\partial K\cap\Omega} \max_{\tau} |[\hat{\epsilon}\frac{\partial U}{\partial n}_{\tau}]|/h_k \text{ on } K \in T.$$

We also state the following analogue of Theorem 2.6 for the SD-method (1.11) for the time dependent problem (1.4), again assuming for simplicity that α and β are constant.

Theorem 2.7. There is a constant C such that

$$(2.9)\qquad \|\hat{u}-U\|_{L_2(Q)} \leq C\|\min(1, \hat{\epsilon}^{-1}h^2)R(U)\|_{L_2(Q)} + \max_{Q_-}\hat{\epsilon}^{\frac{1}{2}},$$

where $Q = \Omega \times I$, with $I = (0,T)$ a given time-interval, $Q_- = (\Gamma_- \times I) \cup (\Omega \times \{0\})$,

$$R(U) = |U_t + \beta\cdot\nabla U + \alpha U - f| + |\mathrm{div}_{h_n}(\hat{\epsilon}\nabla U)| + |\frac{[U]}{k_n}| \text{ on } \Omega \times I_n .$$

Remark 2.12. Note that (2.8) has essentially the form $\|\hat{u}-U\| \leq C\|\min(R(U),1)\|$ which should be compared with the a posteriori estimate for the standard Galerkin method for (1.3) corresponding to choosing $\delta = 0$ and $\hat{\epsilon} = \epsilon$ in (1.10): $\|u - U\| \leq C\,\|R(U)\|$. In a situation with boundary or internal layers $\|R(U)\| \to \infty$ as $h \to 0$ and then adaptive error control is not possible for the standard Galerkin method, cf. Section 4 below.

3. Adaptive algorithms

In this section we present the adaptive algorithms based on the a posteriori error estimates stated above, considering first the elliptic problem (1.1). Starting from the a posteriori enor estimate (2.2) we have the following algorithm for control of $\|D^{2-m}(u-U)\| m = 1,2$: Given an initial triangulation T_0, determine successively triangulations T_j with N_j elements and mesh functions h_j and corresponding approximate solutions U_j, $j = 1,\dots,J$, such that h_j is maximal under the condition

$$(3.1)\qquad \alpha_m\|h_j^m f\|_{L_2(K)} + \beta_m\|h_j^m D^2_{h_{j-1}} U_{j-1}\|_{L_2(K)} \leq \frac{\theta TOL}{\sqrt{N_{j-1}}}, \quad K \in T_{j-1},$$

where J is the smallest integer such that

$$(3.2)\qquad \alpha_m\|h_J^m f\| + \beta_m\|h_J^m D^2_{h_{J-1}} U_j\| \leq TOL.$$

Further, θ is a parameter (here $\theta \sim \frac{1}{2}$), through which we may monitor the total number of steps J required to achieve (3.2). Normally, we may expect to have $J \sim 2$ -

5. Notice that (3.1) seeks to equidistribute the contribution form each element to the global error bound $\alpha_m\|h^2f\| + \beta_m\|D_h^2u\|$.

For the parabolic problem the a posteriori error estimate has the form

$$(3.3) \qquad \|u_N - U_N^-\| \le \max_{n \le N} \mathcal{E}_n(U, h_n, k_n, f),$$

where $\mathcal{E}_n$ is a quantity related to time step n. The adaptive algorithm based on (3.3) for control of $\max_{n \le N} \|u_n - U_n^-\|$ has the following form: For $n = 1,2,...,N$, construct a mesh S_n with N_n elements and mesh function h_n, a time step k_n and corresponding approximate solutions U_n on $\Omega \times I_n$ such that

$$\mathcal{E}_n(U_n, h_n, k_n, f) = TOL,$$

and N_n/k_n is (nearly) minimal. To solve the minimization problem we again seek to equidistribute the contributions from the elements in the space-time discretization of $\Omega \times I_n$. For a precise statement of the adaptive algorithm in this case, see Example 6.2 below.

For the stationary hyperbolic problem (1.3) we may design two adaptive algorithms: (i) one algorithm based on (2.8) and (ii) one algorithm based on (2.8) together with the additional requirement that the mesh is refined until $\hat{\epsilon} = \epsilon$. In the case (i) the adaptive algorithm is obtained by replacing (3.1) by

$$(3.4a) \qquad Ch_j \min(1, \hat{\epsilon}_{j-1}h_{j-1}^2)R(U_{j-1}) \le \frac{\theta TOL}{\sqrt{N_{j-1}}} \quad \text{on } K \in T_{j-1},$$

$$(3.4b) \qquad Ch_j^{\alpha/2}R(U_{j-1})^{1/2} \le TOL \quad \text{if } K \cap \Gamma \ne \phi.$$

In the case (ii) we also add the requirement that

(3.4c) $\quad Ch_j^{\alpha}R(U_{j-1}) \leq \epsilon.$

The stopping criterions are obvious. With proper normalization it appears that $\hat{\epsilon} = \epsilon$ corresponds to resolving all scales of the continuous solution, see Section 4 below.

Extensions to the time dependent hyperbolic problem is obtained by replacing Ω by $\Omega \times I$ and also here we may add the requirement $\hat{\epsilon} = \epsilon$.

Remark 3.1. For error control in the maximum norm ($L_\infty(\Omega)$-norm) e.g. for the Poisson equation, the adaptive algorithm has the form (see [E2]).

$$C(\|h_j^2 f\|_{L_\infty(K)} + \|h_j^2 D^2_{h_{j-1}} U_{j-1}\|_{L_\infty(K)}) = TOL \qquad \text{on } K \in T_{j-1}$$

4. Reliability and efficiency.

We recall that our adaptive algorithms are based on a posteriori error estimates of the form $\|u-U\| \leq \mathcal{E}(U, h, \text{data})$ and that the stopping criterion is $\mathcal{E}(U. h, \text{data}) \leq TOL$, which guarantees that if the stopping criterion is satisfied, then the error will be within the given tolerance and thus the adaptive algorithm is reliable.

Next, we consider the efficiency of our adaptive algorithms. To prove the efficiency in a precise way we need to prove that the final mesh produced by the adaptive algorithm is close to the optimal mesh, which is the mesh with fewest degrees of freedom such that the corresponding approximate solution is within the tolerance. This is possible to show e.g. for the Poisson equation with error control in the maximum norm by using localization techniques to prove that (see [E2]) on a mesh produced by the adaptive algorithm, we have for $x \in \Omega$,

$$\max_{|y-x| \leq Ch} |h^2(y) D^2 u(y)| \geq cTOL$$

This result proves essentially that for all $x \in \Omega$ the local interpolation error is bounded below by a constant times the tolerance and thus that the mesh is now where overly refined. In L_2-norms efficiency in this precise sense is more difficult to prove and in such cases it may be of interest to prove efficiency in a weaker sense. We present a simple result of this type for the Poisson equation, stating that the a posteriori error bounds may be estimatatėd by (sharp) a priori error bounds (see [EJ3]).

Theorem 4.1. Under the assumptions of Theorem 2.1 there is a constant C such that for $m = 1,2$,

$$\alpha_m\|h^m f\| + \beta_m\|h^m D_h^2 U\| \leq C\|h^m D^2 u\|.$$

From this result it follows by Remark 2.2 that in general the global L_2 or H^1-interpolation error on a mesh produced by the adaptive algorithm is not essentially below the given tolerance, which indicates efficiency in a certain sense (but does not necessarily exclude local over-refinement). A somewhat different indication on efficiency also follows from Theorem 4.1, namely that an optimal mesh for which $C\|h^m D^2 u\| =$ TOL, will (up to a constant) be accepted by the stopping criterion of the adaptive algorithm. In particular it follows that it is possible to satisfy the stopping criterion for any tolerance e.g. if $\|D^2 u\|$ is finite.

For the parabolic problem (1.2) one can prove an efficiency result localized in time corresponding to Theorem 4.1 stating that for almost all time steps n the interpolation error E_{on} $(q = 0)$ or E_{2n} $(q = 1)$ is not essentially below the given tolerance on meshes generated by the adaptive algorithm, see [EJ3].

Also for the hyperbolic model problems (1.3) and (1.4) certain results indicating efficiency of our adaptive algorithms are available. Let us give an outline of these results for the SD-method (1.10) for the stationary problem (1.3). Typically, the exact solution u of (1.3) is piecewise smooth with a boundary layer of width $O(\epsilon)$ at the outflow boundary $\Gamma_+ = \Gamma\backslash\Gamma_-$ and internal layers of width $O(\sqrt{\epsilon})$ along streamlines of the velocity field β e.g. if the inflow boundary data is discontinuous In the typical case the continuous solution thus has features on the three different scales $0(1)$, $0(\sqrt{\epsilon})$ and $0(\epsilon)$ in smooth regions, internal layers and outflow layers, respectively. Let us now first consider the adaptive algorithm (3.3) for L_2-norm control based on the a posteriori bound (2.8). In this case theoretical and computational results indicate that the algorithm will produce a mesh with mesh size of order $O(TOL^2)$ at the outflow boundary, $O(TOL^{8/3})$ at an internal layer, and of order $O(TOL)$ in regions where the exact solution is smooth. This follows from localization results showing that the width of the numerical outflow layer is $O(h)$ and the width of the internal numerical layer is $O(h^{3/4})$, (see [JNP], [JSW]) and the fact that the integrand in (2.8) will be of order $O(1)$ in the layers, and by Theorem 2.5 of order $O(h)$ in regions where the exact solution is smooth. Altogether, these results indicate that the algorithm for L_2-norm control will produce a mesh with correct mesh size close to layers and possibly slight over-refinement in smooth regions, since there the a priori error estimate indicates $O(h^{3/2})$ accuracy, while the a posteriori estimate only gives $O(h)$. Notice in particular that the algorithm is able to handle a problem with both boundary and interior layers and smooth parts with a balanced attention to all features. Depending on the tolerance level chosen and ϵ, the algorithm may resolve internal layers (if TOL $\leq 0(\epsilon^{3/16})$) and also outflow layers (if TOL $\leq 0(\epsilon^{1/2})$).

Next, we add an indication to refine if $\hat{\epsilon} > \epsilon$. In an outflow layer we will have $\hat{\epsilon} = 0(h)$ if $\alpha = 2$, and thus $\hat{\epsilon} = \epsilon$ will require $h = 0(\epsilon)$ which corresponds to resolution of the outflow layer of width $0(\epsilon)$. In an internal layer, we will have $\hat{\epsilon} = 0(h^{3/2})$ if $\alpha = 2$, and thus $\hat{\epsilon} = \epsilon$ will require $h = 0(\epsilon^{2/3})$, which again corresponds to resolution of the internal layer of width $0(\epsilon^{1/2})$ since the width of the numerical layer is $0(h^{3/4})$. Of course, the stated results are qualitative in nature and are only valid up to constants, but indicate that with proper normalization the requirement $\hat{\epsilon} = \epsilon$ imposes resolution of all scales of the continuous solution.

5. The structure of the proofs of the a priori and a posteriori error estimates

In this section we briefly outline the structure of the proofs of our a priori and a posteriori error estimates. We start from a continuous problem with the following variational formulation: Find $u \in V$ such that

$$(5.1) \qquad B(u,v) = L(v) \quad \forall v \in V,$$

where $B(..)$ is a continous bilinear form on $V \times V$, L is a continuous linear form on V and V is a Hilbert space (e.g. $H_0^1(\Omega)$ in the case of (1.1) and (1.3)). Next, we consider a corresponding discrete problem: Given a finite element space $V_h \subset V$ find $V \in V_h$ such that

$$(5.2) \qquad B(U,v) = L(v) \quad \forall v \in V_h.$$

To prove an a posteriori error estimate in a norm $\|\cdot\|$ related to the scalar product $(\cdot,\cdot)$, let $\varphi \in V$ be the solution of the continuous dual problem: Find $\varphi \in V$ such that

$$(5.3) \qquad B(w,\varphi) = (w,u-U) \quad \forall w \in V,$$

a problem which we assume to be uniquely solvable. Choosing now $w = u - U$ in (5.3) we have using (5.1)

$$\|u - U\|^2 = B(u - U,\varphi) = B(u,\varphi) - B(U,\varphi) = L(\varphi) - B(U,\varphi),$$

which gives the following error representation formula using (5.2):

$$(5.4) \qquad \|u - U\|^2 = L(\varphi - \tilde{\varphi}) - B(U,\varphi-\tilde{\varphi}),$$

where $\tilde{\varphi} \in V_h$ is an interpolant of φ. The idea is now to establish a stability result for the dual problem (5.3) of the form

$$|||\varphi||| \le C\|u - U\|, \tag{5.5}$$

where the norm $|||\cdot|||$ is as strong as possible, and then estimate $\varphi - \tilde{\phi}$ in a weighted norm (as strong as possible), with weight depending on (a negative power of) the mesh size h, in terms of $C|||\varphi|||$. Inserting this estimate into (5.4) and dividing by $\|u - U\|$ gives an a posteriori error estimate of.the form

$$\|u - U\| \le \mathcal{E}(U,h,L).$$

where $\mathcal{E}(U,h,L)$ depends on U, the mesh size h and the data L. Clearly, the stability estimate (5.5) for the continuous dual problem (5.3) is the critical ingredient; in particular we have to find, analytically or computationally, a reasonable approximation of the best constant in (5.5).

The a priori error estimate is obtained by introducing the following discrete dual problem: Find $\phi \in V_h$ such that

$$B(w,\phi) = (w, \tilde{U} - U),$$

where $\tilde{U} \in V_h$ is an interpolant of u. Choosing here $w = \tilde{U} - U \in V_h$ we get using (5.1) and (5.2)

$$\|\tilde{U} - U\|^2 = B(\tilde{U} - U, \phi) = B(\tilde{U} - u, \phi), \tag{5.6}$$

from which we obtain an estimate for $\|\tilde{U} - U\|$ using a (strong) stability estimate for ϕ again of the form (5.5) (but with a different norm $|||\cdot|||$) and standard interpolation error estimates for $\tilde{U} - u$.

Summing up, the proofs of the a posteriori and a priori error estimates are based on error representation formulas of the form (5.4) and (5.6) together with strong stability estimates for the associated continuous and discrete dual problems, and standard interpolation theory is used to estimate $\varphi - \tilde{\Phi}$ and $u - \tilde{U}$, respectively. The right hand side of (5.4) is clearly related to the <u>residual</u> of the discrete solution U, while the right hand side of (5.6) may be viewed as a <u>truncation</u> <u>error</u>. For the concrete implementation of the above approach, we refer to [E1-J], [EJ1-5], [J1-2].

6. Numerical results

In this section we present the results of some numerical experiments . Here each mesh T_j is obtained from a previous mesh T_{j-1}, starting with a given coarse mesh T_0, by either local refinement dividing certain triangles (fathers) into four similar triangles (sons) by connecting the midpoints of the sides, or local unrefinement replacing a group of four sons by their common father. In particular, the minimal mesh size of the mesh T_j is half of that of T_{j-1}.

<u>Example 6.1</u> We consider the Poisson equation (1.1) on the square $(-1,1)^2$ with $f = 0$ and exact solution $u(x_1,x_2) = \arctan(x_2/(x_1 + 1))$ with nonzero boundary conditions with a discontinuity at (-1,0). In Fig. 6.1 we give the final mesh produced by (3.1) in the case $m = 2$ choosing $\alpha_2 = 0.15$, $\beta_2 = 0.3$ and TOL= 0.01 , together with the estimated and actual L_2-error on the successive meshes.

<u>Example 6.2</u> We consider the following adaptive algorithm for the Discontinuous Galerkin method (1.7) for the parabolic problem (1.2) based on the <u>a</u> <u>posteriori</u> error estimate (2.7): For $n = 1,2,\ldots,N$, given an initial triangulation $T_{n,0}$ and an initial time step $k_{n,0}$, determine successively triangulations $T_{n,j}$ with $N_{n,j}$ elements and mesh functions $h_{n,j}$, time steps $k_{n,j}$ and corresponding approximate solutions $U_{n,j}$ defined on $\Omega \times I_n$, $j = 1,\ldots,J$, such that with $h_{n,j}$ and $k_{n,j}$ maximal

$$C_6 \max_{I_{nj}} \|h_{n,j}^2 f\|_{L_2(K)} + C_8 \max_{I_{nj}} \|h_{n,j}^2 D_{n,j-1}^2 U_{n,j-1}\|_{L_2(K)}$$

$$+ C_{11}\|h_{n,j}^2 [U_{n,j-1}]/k_{n,j-1}\|_{L_2(K)} = \frac{\theta \mathrm{TOL}}{2\sqrt{N_{n,j-1}}} \qquad \forall\, K \in T_{n,j-1},$$

$$k_{n,j}^3 (C_7 \|f_{tt}\|_{I_{n,j}} + \min(C_{10}\|\Delta_{n,j-1} P_{n,j-1}[U_{n,j-1}]_{n-1}/k_{n,j-1}^2\|,$$

$$C_9 \|[U_{n,j-1}]_{n-1}/k_{n,j-1}^3\|)) = \frac{\mathrm{TOL}}{2} \text{ if } q = 1,$$

$$k_{n,j}(C_2\|f\|_{I_{n,j}} + C_4\|[U_{n,j-1}]_{n-1}/k_{n,j-1}\|) = \frac{TOL}{2} \text{ if } q = 0,$$

where $C_2 = 3$, $C_6 = 0.15$, $C_7 = 1/36$, $C_8 = 0.3$, $C_9 = 2$, $C_{10} = 1/6$ and $C_{11} = 0.2$. We choose the initial data u_0 to be an "approximate deltafunction" at $x = 0$: $u_0 = 250 \exp(-|x|^2/250)$, and $\Omega = (0,1)^2$. We give in Fig. 6.2 the sequence of time steps, the number of elements in the triangulation on each time interval, and the $L_2(\Omega)$-error $\|u(t_n) - U_n^-\|$, $n = 1,2,\ldots$, in the case $q = 1$, together with the space mesh at time step 5. We notice that the actual error is approximately constant in time and slightly below the given tolerance.

Example 6.3 We now give some results for the adaptive algorithms for the streamline diffusion method (1.10) for the stationary hyperbolic problem (1.3) based on (i) (3.4a,b) and (ii) (3.4a,b) together with the additional refinement criterion $\hat{\epsilon} = \epsilon$ corresponding to (3.4c). We consider a problem with both internal layer and outflow layer, and with $\Omega = (0,1)^2$, $f = 0$, $\beta = (2,1)$, $\alpha = 0$, $u(0, x_2) = 1$ for $0 < x_2 \leq 1$, $u(x_1,1) = 1$ for $0 \leq x_1 < 1$ and $u(x_1, x_2) = 0$ if $x_1 = 1$ or $x_2 = 0$. In Fig. 6.3 - 6.5 we give some results with the algorithms (i) and (ii) and varying ϵ and TOL. The constants C_i in (1.10) were chosen as follows $C_1 = 0.5$, $C_2 = 0.7$. Note that the width of the layer refinement of mesh T_j is related to the width of the numerical layer of the approximate solution U_{j-1} on mesh T_{j-1}. This is the reason why the width of the refinement of T_j appears to be too large as compared to the width of the numerical layer of the solution U_j. Note also that , for simplicity, the L2-error is computed by comparison with the exact solution corresponding to $\epsilon = 0$, which means that the given L2-error is not precise in the case of refined meshes and relatively large ϵ.

References.

[A] Abdalass, E.M., Resolution performante du probléme de Stokes par mini-element, amillages auto-adaptifs et methods multiquilles-applications. Thése de 3me cycle, Ecole Central de Lyon, 1987.

[B] Babuska, I and Rheinboldt, W. C., Reliable error estimation and mesh adaptation for the finite element method, in Computational Methods in Nonlinear Mechanics, North-Holland,New York (1980), p 67-108.

[Bal] Bank, R., Analysis of a local a posteriori error estimate for elliptic equations, in Accuracy Estimates and Adaptive Refinements in Finite element Computations (Eds. I. Babuska et al.) Wiley, New York, 1986.

[BA2] Bank, R., Personal communication 1989.

[E1] Eriksson, K., Adaptive finite element methods based on optimal error estimates for linear elliptic problems, Technical report, Department of Mathematics, Chalmers Univ. of Techn., 1987.

[E2] Eriksson, K., Adaptive finite element methods for parabolic problems II: A priori error estimates in $L_{\infty}(L_2)$ and $L_{\infty}(L_{\infty})$, Technical report, Department of Mathematics, Chalmers University of Technology, 1988.

[E3] Eriksson, K., Error estimates for the $H_0^1(\Omega)$ and $L_2(\Omega)$ projections onto finite element spaces under weak mesh regularity assumptions, Department of Mathematics, Chalmers University of Technology, 1988.

[EJ1] Eriksson, K. and Johnson, C., An adaptive finite element method for linear elliptic problems, Math. Comp. 50 (1988), pp.361-383.

[EJ2] Eriksson, K. and Johnson, C., Error estimates and automatic time step control for non-linear parabolic problems, I, SIAM J. of Numer. Anal. 24(1987), p.12-23.

[EJ3] Eriksson, K. and Johnson, C., Adaptive finite element methods for parabolic problems I: A linear model problem, to appear in SIAM J. Numer. Anal.

[EJ4] Eriksson, K. and Johnson, C., Adaptive finite element methods for parabolic problems III: Time steps variable in space, IV: A nonlinear problem, to appear.

[EJ5] Eriksson, K. and Johnson C., Adaptive streamline diffusion finite element methods for convection-diffusion problems, Technical report, Dept. of Mathematics, Chalmers Univ. of Technology, 1990.

[EJT] Eriksson, K., Johnson, C. and Thomée, V., Time discretization of parabolic problems by the Discontinuous Galerkin method, RAIRO, MAN 19 (1985), p. 611-643.

[Ew] Ewing, D., Adaptive mesh refinements in large-scale fluid flow simulation, in Accuracy Estimates and Adaptive Refinements in Finite element Computatons (Eds. I. Babuska et al.) Wiley, New York, 1986.

[H] Hansbo, P., Adaptivity and streamline diffusion procedures in the finite element methods, Thesis, Chalmers University of Technology, 1989.

[J] Johnson, C., Error estimates and adaptive time step control for a class of one step methods for stiff ordinary differential equations, SIAM J. of Numer. Anal. 25(1988), pp. 908-926.

[JNP] Johnson, C., Nävert, U. and Pitkäranta, J., Finite element methods for linear hyperbolic problems, Comput. Meth. Appl. Mech. Engrg 45(1984), p 285-312.

[JNT] Johnson, C., Nie, Y.-Y. and Thomée, V., An a posteriori error estimate and automatic time step control for a backward Euler discretization of a parabolic problem, to appear in SIAM J. Numer. Anal.

[JSH] Johnson, C., Szepessy, A., and Hansbo, P., On the convergence of shock-capturing streamline diffusion finite element methods for conservation laws, to appear in Math. Comp.

[JSW] Johnson, C., Schatz, A., and Wahlbin, L., Crosswind smear and pointwise errors in streamline diffusion finite element methods, Math. Comp. 49(1987), p 25-38.

[L] Lennblad, J., An adaptive finite element method for a linear parabolic problem. Technical report, Department of Mathematics, Chalmers University of Technology, 1988.

[Li] Lippold, G., Error estimates and step-size control for the approximate solution of a first order evolution equation, preprint, Akademic der Wissenschaffen der Karl-Weierstrass-institut für Matematik, Berlin, 1988.

[LMZ] Löhner, R., Morgan, K. and Zienkiewicz, Adaptive grid refinement for the compressible Euler equations, in Accuracy Estimates and Adaptive Refinements in Finite element Computations (Eds. I. Babuska et al.) Wiley, New York, 1986.

[ODSD] Oden, J.T., Demkowicz, L., Strouboulis, T. and Devloo, P., Adaptive methods for problems in solid and fluid mechanics, in Accuracy Estimates and Adaptive Refinements in Finite element Computations (Eds. I. Babuska et al.) Wiley, New York, 1986.

[Sz] Szepessy, A., Convergence of the streamline diffusion finite element method for conservation laws, Thesis, Dept. of Mathematics, Chalmers Univ. of Technology, 1989.

[V] Verfürth, R., A posteriori error estimators for the Stokes equations, Numer. Math. 1989.

2 3 7

A POSTERIORI ERROR ESTIMATES FOR THE STOKES EQUATIONS: A COMPARISON

RANDOLPH E. BANK* AND BRUNO D. WELFERT†

Abstract. Several *a posteriori* error estimates for the Stokes equations have been derived by several authors [11] [8]. In this paper we compare some estimates based on the solution of local Stokes systems with estimates based on the residuals of the discretized finite element equations. Their performance as local indicators as well as global estimates is investigated.

Key words. Mixed finite element methods, Stokes equations, a posteriori error estimates, mesh adaptation.

1. Introduction. When numerically solving a set of partial differential equations through a finite element strategy associated with a weak formulation, one usually faces the problem of increasing the accuracy of the solution without adding unnecessary degrees of freedom in non critical parts of the computational domain. In order to identify these regions indicators were created, which allow their automatic determination by computing some function of the characteristic features of the solution, such as indicators based on the gradient of the Mach number in Computational Fluid Dynamics [4] [10], indicators derived from *a priori* error estimates, or indicators involving residuals of the discretized equations [3] [1].

More recently the trend has been to derive *a posteriori* error estimates based on more mathematical criteria, by solving small local problems resembling the original global one, but involving higher order finite elements [5], [6], [11], [7].

In this paper we compare a few of these estimates obtained for the Stokes problem. The finite element scheme used is the classical mini-element formulation, which is recalled in section 2. Three estimates based on the resolution of a local Stokes problem alc ıg with one based only on residuals are presented in section 3. In section 4 a few comparison inequalities are stated, and section 5 examines their numerical behavior on test problems for which an exact solution is known, and on typical examples of CFD as well.

2. The mini-element discretization of the Stokes equations. We consider the classical Stokes problem: Find u (velocity field, 2 components) $\in (\mathcal{H}^1(\Omega))^2$ (the usual Sobolev space) and p (pressure field) $\in \mathcal{L}^2(\Omega)$ (the usual Lebesgue space) such that

$$(1)\qquad \begin{cases} -\nu\Delta u + \nabla p &= f \quad \text{in } \Omega \\ \nabla \cdot u &= 0 \quad \text{in } \Omega \\ u &= g \quad \text{on } \partial\Omega \end{cases}$$

* Department of Mathematics, University of California at San Diego, La Jolla, California 92093. The work of this author was supported by the Office of Naval Research under contract N00014-89J-1440, by Avions Marcel Dassault - Breguet Aviation, 78 quai Marcel Dassault, 92214 St Cloud, France and by Direction des Recherches Etudes et Techniques, 26 boulevard Victor, 75996 Paris Armées, France.

† Department of Mathematics, University of California at San Diego, La Jolla, California 92093. The work of this author was supported by Avions Marcel Dassault - Breguet Aviation, 78 quai Marcel Dassault, 92214 St Cloud, France and by Direction des Recherches Etudes et Techniques, 26 boulevard Victor, 75996 Paris Armées, France.

in a bounded domain $\Omega \subset \mathcal{R}^2$ (ν is a viscosity parameter).

A weak formulation of equations (1) can be derived using integration by parts, and can be shown to satisfy an LBB condition, thus providing a unique solution to the resulting system [8] (up to an arbitrary additive constant for the pressure).

Let $\mathcal{T}$ denote a triangulation of the domain Ω, such that any two triangles in $\mathcal{T}$ share at most a vertex or an edge. Let h_τ be the diameter of a triangle $\tau \in \mathcal{T}$ and $h = \max\limits_{\tau \in \mathcal{T}} h_\tau$. E is the set of interior edges. For $e \in E$, h_e denotes the length of e. We suppose also that the triangulation $\mathcal{T}$ satisfies a minimal angle condition, i.e. the smallest angle in triangle $\tau \in \mathcal{T}$ is bounded away from zero by some constant independent of h. This implies in particular that $C_1 h_\tau \le h_e \le C_2 h_\tau$ for $e \in \partial\tau$. Furthermore, for $\Gamma = e$, E, or some subset of E, we define the inner product

$$< u, v >_\Gamma = \int_\Gamma uvds = \sum_{e \in \Gamma} \int_e uvds.$$

Let $\mathcal{C}^0$ be the space of continuous functions over $\mathcal{T}$. Let $\psi_i = \psi_i(\tau), i = 1, 3$ be the barycentric coordinates (linear nodal basis functions) in the triangle τ. Then we introduce the spaces

$$\begin{aligned}
&(2) && \mathcal{H}_\mathcal{T} &=& \prod_{\tau \in \mathcal{T}} \mathcal{H}^1(\tau) = \{u, u_{|\tau} \in \mathcal{H}^1(\tau), \tau \in \mathcal{T}\} \\
&(3) && \mathcal{L} &=& \prod_{\tau \in \mathcal{T}} \mathcal{L}_\tau = \prod_{\tau \in \mathcal{T}} span\{\psi_i(\tau), 1 \le i \le 3, \tau \in \mathcal{T}\} \\
&(4) && \mathcal{K} &=& \prod_{\tau \in \mathcal{T}} \mathcal{K}_\tau = \prod_{\tau \in \mathcal{T}} span\{\psi_i(\tau)\psi_j(\tau), 1 \le i < j \le 3, \tau \in \mathcal{T}\} \\
&(5) && \mathcal{B} &=& \prod_{\tau \in \mathcal{T}} \mathcal{B}_\tau = \prod_{\tau \in \mathcal{T}} span\{\psi_1(\tau)\psi_2(\tau)\psi_3(\tau), \tau \in \mathcal{T}\} \\
&(6) && \mathcal{X} &=& (\mathcal{H}_\mathcal{T} \cap \mathcal{L} \cap \mathcal{C}^0 \oplus \mathcal{B})^2 \\
&(7) && \mathcal{Y} &=& \mathcal{L}_0^2 \cap \mathcal{L} \cap \mathcal{C}^0
\end{aligned}$$

and set $\mathcal{Q}_\tau = \mathcal{L}_\tau \oplus \mathcal{K}_\tau$ for $\tau \in \mathcal{T}$ and $\mathcal{Q} = \prod_{\tau \in \mathcal{T}} \mathcal{Q}_\tau$. $\mathcal{L}$ is the space of piecewise linear functions and $\mathcal{Q}$ the space of piecewise quadratic functions on $\mathcal{T}$. The elements of $\mathcal{B}$ are *bubble functions* which vanish on all edges of the triangulation.

The *mini-element* discretization [2] of the weak equations is then given by:

Find $(u_h, p_h) \in \mathcal{X}_g \times \mathcal{Y}$ such that for all $(v, q) \in \mathcal{X}_0 \times \mathcal{Y}$:

$$(8) \qquad \left\{ \begin{array}{lcrcl} a(u_h, v) & + & b(v, p_h) & = & (f, v) \\ b(u_h, q) & & & = & 0 \end{array} \right.$$

This formulation also satisfies an inf-sup condition, which implies the unique solvability of the system (8).

The decomposition $u_h = u_{h,l} + u_{h,b}$, with $u_{h,l} \in (\mathcal{L} \cap \mathcal{C}^0)^2$ and $u_{h,b} \in \mathcal{B}^2$, is unique. In fact, $u_{h,l}$ is usually a better approximation to u than u_h itself (see [11]) and is therefore used in most *a posteriori* estimates using the mini-element formulation, through the residuals r and s defined as

$$(9) \qquad \left\{ \begin{array}{rcl} r & = & f - \nabla p_h \\ s & = & -\nabla \cdot u_{h,l} \end{array} \right.$$

Likewise we define the error terms $e \equiv u - u_{h,l}$ (instead of $u - u_h$) and $\epsilon \equiv p - p_h$. We introduce the (nonsymmetric) bilinear form

$$D((u,p),(v,q)) = \nu(\nabla u, \nabla v) - (p, \nabla \cdot v) + (q, \nabla \cdot u) + \sum_{\tau \in \mathcal{T}} \frac{1}{3600\sigma_\tau \nu}(\nabla p, \nabla q)_\tau$$

and the semi-norm N_i, $i = 1,2,3$, defined by

$$\begin{aligned} N_1(u,p)^2 &\equiv \nu\|\nabla u\|^2 + \frac{\|p\|^2}{\nu} \\ N_2(u,p)^2 &\equiv \nu\|\nabla u\|^2 + \sum_{\tau \in \mathcal{T}} \frac{1}{3600\sigma_\tau \nu}\|\nabla p\|_\tau^2 \\ N_3(u,p) &\equiv \sqrt{\nu}\|\nabla u\| + \frac{\|p\|}{\sqrt{\nu}} \end{aligned}$$

Here $\sigma_\tau = \frac{1}{|\tau|}(\nabla\psi_b(\tau), \nabla\psi_b(\tau))_\tau$ ($|\tau|$ represents the area of the triangle τ) and $\psi_b(\tau) = \psi_1(\tau)\psi_2(\tau)\psi_3(\tau)$ in triangle τ (bubble function). σ_τ is of order $\mathcal{O}(h_\tau^{-2})$ in the sense that there exist two positive constants C_3 and C_4 depending on the minimal angle in the triangulation such that

$$C_3\, h_\tau^2 \le \sigma_\tau^{-1} \le C_4\, h_\tau^2$$

Note that on $(\mathcal{K} \oplus \mathcal{B}) \times (\mathcal{K} \oplus \mathcal{B}) \times \mathcal{K}$, N_i, $i = 1,2,3$, define equivalent norms, i.e.

$$\beta_1 N_2(u,p) \le N_1(u,p) \le N_3(u,p) \le \beta_2 N_2(u,p) \qquad (u,p) \in (\mathcal{K} \oplus \mathcal{B}) \times (\mathcal{K} \oplus \mathcal{B}) \times \mathcal{K}$$

for some positive constants β_1 and β_2.

3. Estimates for the Stokes problem. In this section we introduce three different estimates/indicators for the problem presented in section 2. Two of them are based on some norm N_i of the solution of small local Stokes (or modified Stokes) systems, and the remaining one simply uses the residuals of the discretized equations; $\left[\frac{\partial u_{h,l}}{\partial n}\right]_J$ denotes the jump of $\frac{\partial u_{h,l}}{\partial n}$ across an edge $e \in \partial\tau \cap \Omega$:

- In [8] an estimate η_1 was derived from an error analysis performed on a Petrov-Galerkin formulation of the Stokes equations. Errors in both velocity and pressure components were approximated by elements in $\mathcal{K}$ (quadratic bump functions), so that the following 9×9 (modified) Stokes system was solved in each triangle: in triangle τ find $(e'_\tau, \epsilon'_\tau) \in \mathcal{K}_\tau \times \mathcal{K}_\tau \times \mathcal{K}_\tau$ such that:

$$\left\{ \begin{aligned} \nu(\nabla e'_\tau, \nabla v)_\tau - (\epsilon'_\tau, \nabla \cdot v)_\tau &= (r, v)_\tau + \frac{\nu}{2} < \left[\frac{\partial u_{h,l}}{\partial n}\right]_J, v >_{\partial\tau\cap\Omega} \\ (q, \nabla \cdot e'_\tau)_\tau + \frac{(\nabla\epsilon'_\tau, \nabla q)_\tau}{3600\sigma_\tau \nu} &= (s, q)_\tau + \frac{1}{3600\sigma_\tau \nu}(r, \nabla q)_\tau \end{aligned} \right.$$

for all $(v,q) \in \mathcal{K}_\tau \times \mathcal{K}_\tau \times \mathcal{K}_\tau$.
Then $\eta_{1,\tau}^2 \equiv N_i(e'_\tau, \epsilon'_\tau)$, $i = 1,2$ or 3, and we define

$$\eta_1^2 \equiv \sum_{\tau \in \mathcal{T}} \eta_{1,\tau}^2 = N_i(e', \epsilon')$$

Remark: the term $(\nabla\epsilon'_\tau, \nabla q)_\tau$ is a stabilization term. Indeed, had we not introduced this term (and hence the corresponding term $(r, \nabla q)_\tau$ in the right-hand side), the resulting 9×9 system would then have been possibly singular (i.e. not satisfying a local LBB condition), in particular for boundary triangles, where the number of degrees of freedom in the pressure error would then be too large compared to the number of degrees of freedom for the non-Dirichlet velocity unknowns. However, for interior triangles, an estimate without the stabilization terms can be shown to be well defined, and is equivalent to $\eta_{1,\tau}$.

- Alternatively, R. Verfürth analyzed an error estimate based on the solution of an 11 by 11 Stokes system in each triangle [11]: in triangle τ find $(e''_\tau, \epsilon''_\tau) \in (\mathcal{K}_\tau \oplus \mathcal{B}_\tau) \times (\mathcal{K}_\tau \oplus \mathcal{B}_\tau) \times \mathcal{K}_\tau$ such that:

$$\begin{cases} \nu(\nabla e''_\tau, \nabla v)_\tau - (\epsilon''_\tau, \nabla \cdot v)_\tau &= (r, v)_\tau + \frac{\nu}{2} < \left[\frac{\partial u_{h,l}}{\partial n}\right]_J, v >_{\partial\tau\cap\Omega} \\ (q, \nabla \cdot e''_\tau)_\tau &= (s, q)_\tau \end{cases}$$

for all $(v, q) \in (\mathcal{K}_\tau \oplus \mathcal{B}_\tau) \times (\mathcal{K}_\tau \oplus \mathcal{B}_\tau) \times \mathcal{K}_\tau$.
Then put $\eta^2_{2,\tau} \equiv \mathcal{N}_i(e''_\tau, \epsilon''_\tau)$, $i = 1, 2$ or 3, and define

$$\eta_2^2 \equiv \sum_{\tau\in\mathcal{T}} \eta^2_{2,\tau} = \mathcal{N}_i(e'', \epsilon'')$$

Since both velocity components of the solution of the previous system decomposes uniquely onto $\mathcal{K}_\tau \oplus \mathcal{B}_\tau$ as $e''_\tau = e''_{\tau,q} + e''_{\tau,b}$, one may compute for the estimate $\eta^{(i)2}_{2,\tau} \equiv \mathcal{N}_i(e''_{\tau,q}, \epsilon''_\tau)$ and set

$$\eta_2^{(i)2} \equiv \sum_{\tau\in\mathcal{T}} \eta^{(i)2}_{2,\tau}$$

- Finally, we define the a posteriori error estimate η_3 in triangle τ by computing

$$\eta^2_{3,\tau} \equiv |\tau|\frac{||r||^2_\tau}{\nu} + \nu||s||^2_\tau + \frac{\nu}{2}\sum_{e\in\partial\tau\cap\Omega} h_e \left\| \left[\frac{\partial u_{h,l}}{\partial n}\right]_J \right\|^2_e$$

in all triangles and setting $\eta_3^2 \equiv \sum_{\tau\in\mathcal{T}} \eta^2_{3,\tau}$. A slightly different form of this estimate was presented in [11].

In the next section we compare these estimates with each other and with the exact discretization error, locally as well as globally.

4. Comparison of the estimates. In this section we state local and global comparison results between the three estimates introduced in section 3. We will use the following inequalities, for $v \in \mathcal{K}$:

$$\begin{aligned} &(10) & ||v||_\tau &\leq c_1|\tau|^{1/2}\,||\nabla v||_\tau \\ &(11) & ||v||_e &\leq c'_3 h_e^{-1/2}||v||_\tau \leq c_3 h_e^{1/2}\,||\nabla v||_\tau \\ &(12) & ||\nabla v||_\tau &\leq c'_4 h_\tau^{-1}\,||v||_\tau \leq 3600 c_4 \sigma_\tau |\tau|^{1/2} ||v||_\tau \end{aligned}$$

LEMMA 4.1. *There exist a constant C depending on the minimal angle in the triangulation only such that*

$$D((e'_\tau, \epsilon'_\tau), (v, q)) \leq C\, \eta_{3,\tau}\, N_1(v, q)$$

for all $(v, q) \in \mathcal{K}_\tau \times \mathcal{K}_\tau \times \mathcal{K}_\tau$.

Proof. Using the definition of the error estimate (e', ϵ') and inequalities (10), (11) and (12) we have

$$\begin{aligned} D((e'_\tau, \epsilon'_\tau); (v, q)) \;&=\; (r, v)_\tau + \nu < \left[\frac{\partial u_{h,l}}{\partial n}\right]_J, [v]_A >_{\partial\tau\cap\Omega} +(s, q)_\tau + \frac{(r, \nabla q)_\tau}{3600\sigma_\tau\nu} \\ &\leq \|r\|_\tau\|v\|_\tau + \frac{\nu}{2}\sum_{e\in\partial\tau\cap\Omega}\left\|\left[\frac{\partial u_{h,l}}{\partial n}\right]_J\right\|_e \|v\|_e + \|s\|_\tau\|q\|_\tau + \|r\|_\tau\frac{\|\nabla q\|_\tau}{3600\sigma_\tau\nu} \\ &\leq c_1|\tau|^{1/2}\|r\|_\tau\|\nabla v\|_\tau + \frac{\nu}{2}\sum_{e\in\partial\tau\cap\Omega} c_3 h_e^{1/2}\left\|\left[\frac{\partial u_{h,l}}{\partial n}\right]_J\right\|_e \|\nabla v\|_\tau \\ &\quad +\|s\|_\tau\,\|q\|_\tau + c_4\frac{|\tau|^{1/2}}{\nu}\|r\|_\tau\,\|q\|_\tau \\ &\leq \sqrt{2}\left(\frac{c_1^2|\tau|}{\nu}\|r\|_\tau^2 + \nu\frac{c_3^2}{4}\sum_{e\in\partial\tau\cap\Omega} h_e\left\|\left[\frac{\partial u_{h,l}}{\partial n}\right]_J\right\|_e^2 + \nu\|s\|_\tau^2 + c_4^2\frac{|\tau|}{\nu}\|r\|_\tau^2\right)^{1/2} \\ &\quad \cdot\left(\nu\|\nabla v\|_\tau^2 + \frac{\|q\|_\tau^2}{\nu}\right)^{1/2} \\ &\leq CN_1(v, q)\,\eta_{3,\tau} \end{aligned}$$

□

By applying this lemma to the quadratic bump functions $(v, q) = (e'_\tau, \epsilon'_\tau)$ we then get a local comparison inequality between the estimates $\eta_{1,\tau}$ and $\eta_{3,\tau}$, namely:

THEOREM 4.2. *In any triangle $\tau \in \mathcal{T}$ we have*

$$\eta_{1,\tau} \leq C\, \eta_{3,\tau}$$

for some constant C depending on the minimal angle in the triangulation.

Proof. A direct application of lemma 4.1 in triangle τ gives

$$\beta_1^2 N_1(e'_\tau, \epsilon'_\tau)^2 \leq N_2(e'_\tau, \epsilon'_\tau)^2 = D((e'_\tau, \epsilon'_\tau), (e'_\tau, \epsilon'_\tau)) \leq C\, N_1(e'_\tau, \epsilon'_\tau)\, \eta_{3,\tau}$$

and the equivalence of the three norms N_1, N_2 and N_3 on $\mathcal{K} \times \mathcal{K} \times \mathcal{K}$ proves the theorem. □

COROLLARY 4.3. *There exist a constant C depending only on the minimal angle in the triangulation such that*

$$\eta_1 \leq C\, \eta_3$$

THEOREM 4.4. *There exist constants C_5, C_6 and C_7 such that*

$$C_5\, \eta_{3,\tau} - C_6\, \|f - P_0 f\|_\tau \leq \eta_{2,\tau} \leq C_7\, \eta_{3,\tau}$$

Proof. see [11]. □

THEOREM 4.5. *There exists a constants C_8 depending only on the minimal angle in the triangulation such that*

$$N_3(e,\epsilon) \le C_3\,\eta_3$$

Also there exist constants C_9, C_{10}, C_{11} and C_{12} such that

$$C_9\,\eta_1 - C_{10}h^2 \le N_1(e,\epsilon) \le C_{11}\,\eta_1 + C_{12}h^2$$

Proof. see [11], [8]. □

In the next section we verify these comparison inequalities on several numerical examples, and test the comparative efficiency of the different estimates.

5. Numerical results. In this section we test both the local efficiency of the different estimates as indicators used to control an automatic mesh adaptation process, and their global behaviour as approximations of the discretization error, on several numerical examples; since we are interested in particular in computing the "effectivity" ratio q (ratio of the estimated global error to the actual global error), and the convergence rates on the different sequences of meshes produced by adaptive refinement, two test problems for which an exact solution is known are considered first. These are a good indication of the general trend and provide a useful insight on the comparative advantages/drawbacks of the estimates tested, especially on problems with singularities arising in the solution of the numerical pde's due to discontinuities in the boundary conditions or in the smoothness of the geometry, as is the case in problems encountered in Computational Fluid Dynamics (Driven Cavity, Backward Facing Step,...).

5.1. Disk with a Crack. We first test our error estimate on a Stokes flow in a disk of radius 1 with a crack joining the center to the boundary; the right-hand side $\boldsymbol{f}$ is 0 and the boundary conditions are

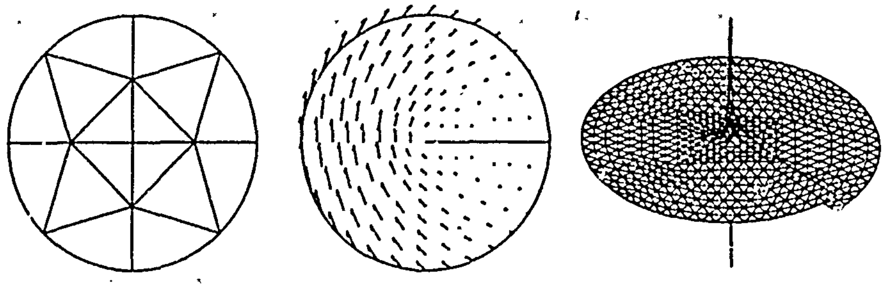
$$\boldsymbol{u}_b^t = \left(\cos\frac{\theta}{2} - \cos\frac{3\theta}{2},\ 3\sin\frac{\theta}{2} - \sin\frac{3\theta}{2}\right)$$

where (r,θ) is a polar representation of a point in the disk. The exact solution is then given by

$$\boldsymbol{u} = \sqrt{r}\,\boldsymbol{u}_b \quad\text{and}\quad p = -\frac{4}{\sqrt{r}}\cos\frac{\theta}{2}$$

and is singular at the crack tip. For this example, $(\nu = 1)$.

We first solve our Stokes problem on a coarse grid (Figure 1(a)), then refine either uniformly or adaptively, thus creating two sequences of meshes of increasing and comparable size. Effectivity ratios, defined as the ratios between the estimated norm of the error (using either one of the norms N_i) to the norm of the exact discretization error, are given for all intermediate meshes in both uniform and adaptive sequences, and allow us to compare the relative efficiency of the η_i as global estimates (see Tables 1 and 2).



(a) (b) (c)

FIG. 1. (a) *initial triangulation* ; (b) *velocity* ; (c) *pressure.*

Disk - Uniform refinement											
NV	NT	q_1			q_2			$q_2^{(i)}$			$q_3^{(*)}$
		N_1	N_2	N_3	N_1	N_2	N_3	N_1	N_2	N_3	
15	16	0.74	1.68	0.81	0.98	2.24	1.05	1.38	3.22	1.39	1.67
45	64	0.99	1.39	0.99	1.30	1.83	1.25	1.81	2.63	1.65	2.28
153	256	1.10	1.35	1.09	1.45	1.77	1.37	2.03	2.54	1.81	2.57
561	1024	1.13	1.32	1.11	1.49	1.73	1.40	2.07	2.48	1.84	2.66
2145	4096	1.14	1.30	1.12	1.49	1.70	1.40	2.08	2.43	1.85	2.69

$^{(*)}$ using norm N_1

TABLE 1
Global effectivity ratios for a sequence of uniform meshes.

The ratio q_3 is computed using the norm N_1 for the discretization error; the use of other norms is not relevant since η_3 is not defined by a norm. A comparison with q_1 shows also that the corollary 4.3 is satisfied with a c· ·nstant $C \leq 1$. All estimates based on the solution of a local system seem to perform equally well on adaptive grids and on uniform meshes, and the ratios q_1, q_2 and $q_2^{(i)}$ remain close to 1.00 (asymptotically exact estimates). As expected the results for the estimate η_3 based only on the norm of the residuals are not as accurate as those using the estimates based on the solution of local systems, especially when applied to highly non uniform meshes, as it is the case for the adaptive strategy. Alltogether, the estimate η_1 associated with the norm N_1 or N_3 seem to approximate globally better the discretization error.

Its local behaviour compares also favorably with the others, in the sense that it ultimately produces a higher convergence rate for the solution, as can be seen in Table 3, even though all estimates appear as good indicators to be used in an adaptive remeshing strategy, yielding grids which aren't very different one from each other (Figure 2). η_1 seems to be also more "asymptotically exact" than the other estimates. Note that the convergence results are in agreement with [9]. Also, the meshes produced may not be exactly symmetrical with respect to the axes of symmetry of the problem (in this case the horizontal axis) (see 2(c)). This is only due to the refinement procedure, according to which the grids are progressively refined to an increasing target number of nodes. Hence a triangle might be refined while its symmetric counterpart is not in order to achieve a particular target value in NV. Further refinement tends to resolve this non-symmetry, but does not necessarily lead to an exactly symmetric mesh again.

5.2. $\frac{3\pi}{2}$-angular sector. The second test problem is a free Stokes flow in a circular sector of angle $3\pi/2$; boundary conditions are

$$\begin{aligned} u_b^t \;&=\; \frac{1}{2}\{(\xi^2-1)[\cos(2-\alpha)\theta - \cos\alpha\theta - \xi^{-1}\sin(2-\alpha)\theta] - \xi^{-1}(1+3\xi^2)\sin\alpha\theta, \\ &\quad (\xi^2-1)[\xi^{-1}(\cos(2-\alpha)\theta - \cos\alpha\theta) + \sin(2-\alpha)\theta] - (\xi^2+3)\sin\alpha\theta\} \end{aligned}$$

where we have $\alpha = \frac{856399}{1572864} \simeq 0.54$ and $\xi = \sqrt{\frac{1+\alpha}{1-\alpha}} \simeq 1.84$. The exact solution is given by

$$u = r^\alpha\, u_b \quad \text{and} \quad p = 2r^{-(1-\alpha)}(\xi^2-1)\{\cos(1-\alpha)\theta - \xi^{-1}\ \sin(1-\alpha)\theta\}$$

and is also singular at the center of the disk, although the degree of singularity is now lower than in the previous example ($\alpha > 0.50$).

Tables 4 and 5 are identical to tables 1 and 2 and confirm the nice features of the estimate η_1.

Note that the convergence rate for this problem is slightly higher than for the previous problem, according to the fact that the solution is "less singular" at the origin.

Here convergence rates are almost identical, which is a consequence of an similar ability from the estimates to resolve the spike in the pressure at the reentrant corner, by producing comparable levels of refinement at that point (see Figures 4(a)(b)(c)(d)).

5.3. A smooth solution in a square. In order to test the local behaviour of the error estimates (we consider here only the estimate η_1 since all indicators lead to similar meshes), we solve the Stokes problem 1 in a square $[0,\pi/2]\times[0,\pi/2]$, with the loading term f and the boundary condition g defined by:

$$f^t = (0, -4\cos x\sin y) \quad \text{and} \quad g^t = \begin{cases} (0,-\sin y) & \text{if} \quad x=0 \\ (\cos y, 0) & \text{if} \quad x=\pi/2 \\ (\sin x, 0) & \text{if} \quad y=0 \\ (0,-\cos x) & \text{if} \quad y=\pi/2 \end{cases} \tag{13}$$

The exact solution to this problem is the smooth function $u^t = (\sin x\cos y, -\cos x\sin y)$ and $p = 2\cos x\cos y - 8/\pi^2$ (so that p has average value 0). Figures 5(a)(b)(c) show that a progressive adaptation of the mesh to the solution using the estimate η_1 leads correctly to almost uniform grids, revealing that no particularly sensitive area was discovered by the estimate.

We now turn to two classical examples in Computational Fluid Dynamics (driven cavity problem and backward facing step problem) which are good test problems to determine the potential of the different estimates presented in this paper.

Disk - Adaptive refinement										
$NV^{(*)}$	$NT^{(*)}$	q_1			q_2			$q_2^{(i)}$		$q_3^{(**)}$
		N_1	N_2	N_3	N_1	N_2	N_3	N_1	N_2	N_3	
15	16	0.74	1.68	0.81	0.98	2.24	1.05	1.38	3.22	1.39	1.67
42	64	0.50	1.30	0.54	1.08	2.23	1.12	1.44	3.13	1.43	1.98
151	267	0.73	1.37	0.73	1.02	1.96	1.02	1.07	2.52	1.05	1.86
558	1045	0.93	1.32	0.88	1.22	1.80	1.13	1.40	2.45	1.28	2.73
2140	4142	1.11	1.29	1.03	1.31	1.62	1.19	1.70	2.25	1.51	3.38

(*) average number of nodes (resp. triangles) obtained when refining the initial mesh using the different estimates η_1, η_2 or $\eta_2^{(i)}$ combined with the norms N_1, N_2 or N_3

(**) using norm N_1

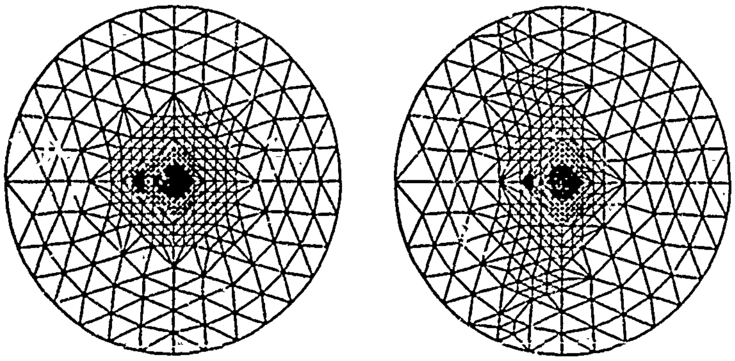
TABLE 2

Global effectivity ratios for a sequence of adapted meshes.

Disk - Convergence Rate					
	uniform	η_1	η_2	$\eta_2^{(i)}$	η_3
N_1	0.87	1.46	1.38	1.31	1.42
N_2	0.60	1.11	1.10	1.07	1.11
N_3	0.83	1.41	1.33	1.26	1.37

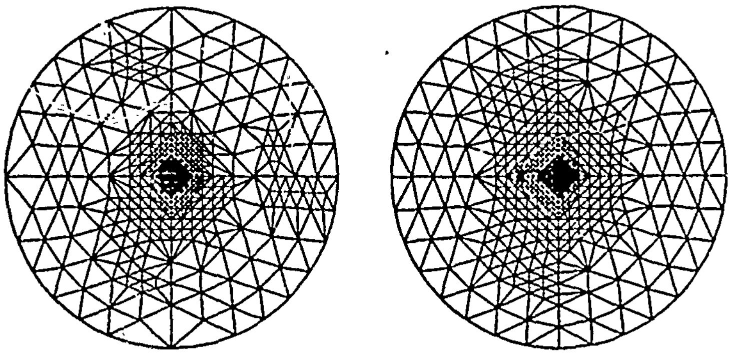
TABLE 3

Global rate of convergence for different adaptive vs uniform strategies.



(a)

(b)



(c) (d)

FIG. 2. *Adapted grids using estimates* (a) η_1 ; (b) η_2 ; (c) $\eta_2^{(i)}$; (d) η_3 *($NV \simeq 558$).*

Sector - Uniform refinement											
NV	NT	q_1			q_2			$q_2^{(i)}$			$q_3^{(*)}$
		N_1	N_2	N_3	N_1	N_2	N_3	N_1	N_2	N_3	
12	12	0.69	1.74	0.76	0.91	2.27	1.00	1.25	3.23	1.30	1.63
35	48	0.93	1.37	0.91	1.16	1.71	1.10	1.57	2.36	1.42	2.34
117	192	1.07	1.31	1.03	1.34	1.63	1.24	1.82	2.26	1.60	2.72
425	768	1.12	1.29	1.09	1.40	1.60	1.29	1.89	2.20	1.66	2.86
1617	3072	1.13	1.27	1.08	1.40	1.57	1.30	1.90	2.16	1.67	2.91

(*) using norm N_1

TABLE 4
Global effectivity ratios for a sequence of uniform meshes.

Sector - Adaptive refinement											
$NV^{(*)}$	$NT^{(*)}$	q_1			q_2			$q_2^{(i)}$			$q_3^{(**)}$
		N_1	N_2	N_3	N_1	N_2	N_3	N_1	N_2	N_3	
12	12	0.69	1.74	0.76	0.91	2.27	1.00	1.25	3.23	1.30	1.63
37	55	0.73	1.38	0.73	1.35	2.27	1.34	1.75	3.06	1.66	2.08
114	195	0.96	1.57	0.93	0.91	1.84	0.89	1.36	2.82	1.27	2.34
427	789	1.06	1.41	0.99	1.53	1.98	1.43	1.68	2.50	1.48	2.93
1618	3092	1.21	1.35	1.14	1.56	1.76	1.44	1.86	2.13	1.65	3.57

(*) average number of nodes (resp. triangles) obtained when refining the initial mesh using the different estimates η_1, η_2 or $\eta_2^{(i)}$ combined with the norms N_1, N_2 or N_3
(**) using norm N_1.

TABLE 5
Global effectivity ratios for a sequence of adapted meshes.

5.4. Driven cavity problem. The system of equations (1) is solved in a unit square, with a loading factor $f^t = (0,0)$ and a unit tangential boundary velocity $g^t = (1,0)$ on the top side and $(0,0)$ elsewhere ($\nu = 0.01$) The domain is initially (symmetrically) triangulated into 8 triangles such that no triangle has 2 boundary edges (Figure 6(a)), and then refined up to 250 nodes using the estimates η_1, $\eta_2^{(i)}$, η_2 and η_3 (Figures 7(a),(b),(c),(d) respectively). The grids obtained are quite identical; we can however point out that the estimates η_1 and η_3 produce more regular meshes (especially at the square's center), even though all of them seem to nicely resolve the two discontinuities in the top corners. Note that the meshes are symmetric, because the refinement process was started with the symmetric coarse grid 6(a). If instead we use the initial grid 6(b) and refine it to $NV \simeq 250$ with η_1, then the balance between the two sides is not kept (Fig 8(a)), thus yielding a pressure level much greater on one side (right, or positive pressure) than on the other. A first remedy could be to refine at a slower pace, in particular by computing estimates based on solutions calculated on intermediate meshes, thus reducing the negative effects of the interpolation procedure (Fig 8(b)). Finally, the initial grid 6(c), with two upper corner triangles having two boundary edges, leads to the mesh 8(c). The refinement is then again quasi-symmetric, although it appears to be more localized than on 7(a).

5.5. Backward facing step problem. As a final example we solve the Stokes equations on a step (Figure 9(a) with a null loading factor f and zero boundary conditions everywhere except on both left and right sides where parabolic inflow and outflow conditions respectively are prescribed, so that the flux is conserved throughout the domain. Then, again, it appears from figures 9(b)(c)(d)(e) that the estimates η_1 and η_3 yield more acceptable meshes than η_2 or $\eta_2^{(i)}$.

6. Conclusion. All estimates seem to be good indicators of the regions in the grids requiring refinement and produce quite similar meshes. Although more expensive to compute than the estimate η_3 based only on the residuals (about two times cheaper than η_1, and slightly more than two times than η_2 and $\eta_2^{(i)}$), the estimates η_1 and η_2 give a more accurate indication of the global error and represent only about a fourth of the computing time needed for the solution process. We also note that the meshes produced by the adaptive strategy have local errors in average greater than those obtained on uniform meshes, but also more uniformly distributed. The estimate η_2, where the total norm (including the bubble term) is considered, appears to give better results than $\eta_2^{(i)}$, however not as good as η_1.

Finally we can point out that this estimate was obtained by computing some norm of an error vector (either 9 × 1 or 11 × 1), thus condensing all directional indications, which could be possibly contained in those vectors, into a single number. Instead, one could use all components of these vectors to create an error estimate that would take the directional errors into account, e.g. along the principal directions of the flow (boundary layers, uni-directional phenomena,...).

REFERENCES

[1] E. M. ABDALASS, *Resolution performante du probleme de stokes par mini-elements, maillage adaptatifs et methodes multigrilles-applications.*, tech. report, These de 3eme cycle, Ecole Centrale de Lyon, 1987.

[2] D. N. ARNOLD, F. BREZZI, AND M. FORTIN, *A stable finite element for the Stokes equations*, Calcolo, 21 (1984), pp. 337–344.

[3] I. BABUSŠKA AND W. C. RHEINBOLDT, *A posteriori error estimates for the finite element method*, Int. J. Numer. Methods Eng., 12 (1978), pp. 1597–1615.

[4] I. BABUŠKA, O. C. ZIENKIEWICZ, J. GAGO, AND E. R. DE A. OLIVEIRA, *Accuracy Estimates and Adaptive Refinements in Finite Element Computations*, John Wiley and Sons, 1986.

[5] R. E. BANK, *Analysis of a local a posteriori error estimate for elliptic equations*, in Accuracy Estimates and Adaptive Refinements in Finite Element Computations, John Wiley and Sons, 1986, pp. 119–128.

[6] R. E. BANK AND A. WEISER, *Some a posteriori error estimators for partial differential equations*, Math. of Comp., 44 (April 1985), pp. 283–301.

[7] R. E. BANK AND B. D. WELFERT, *A posteriori error estimates for the Stokes problem*, (to appear).

[8] R. E. BANK, B. D. WELFERT, AND H. YSERENTANT, *A class of iterative methods for solving saddle point problems*, Numer. Math., (to appear).

[9] F. BREZZI AND J. J. DOUGLAS, *Stabilized mixed methods for the Stokes problem*, Numer. Math., 53 (1988), pp. 225–235.

[10] R. GLOWINSKITH AND J. F. PERIAUX, *Numerical methods for nonlinear problems in fluid dynamics*, tech. report, Proceeding of the International Seminar on Scientific Supercomputers, February 1987.

[11] R. VERFÜRTH, *A posteriori error estimators for the Stokes equations*, Numer. Math., 55 (1989), pp. 309–325.

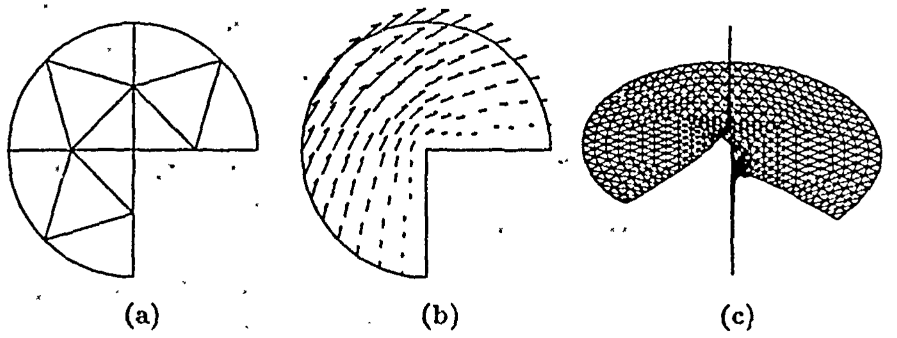
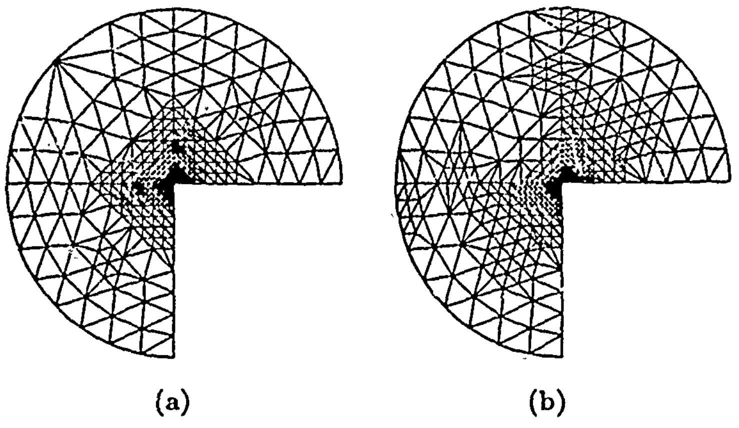


FIG. 3. (a) *initial triangulation* ; (b) *velocity* ; (c) *pressure.*

Sector - Convergence Rate					
	uniform	η_1	η_2	$\eta_2^{(i)}$	η_3
N_1	0.99	1.46	1.46	1.45	1.49
N_2	0.67	1.11	1.13	1.13	1.16
N_3	0.94	1.42	1.42	1.40	1.45

TABLE 6
Global rate of convergence for different adaptive vs uniform strategies.



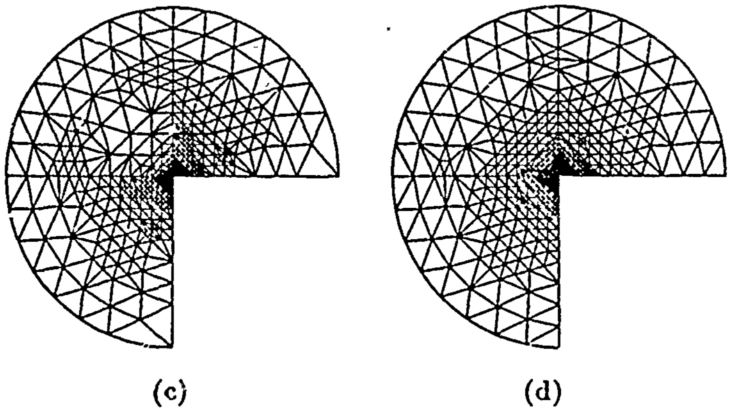


FIG. 4. *Adapted grids using estimates* (a) η_1 *;* (b) η_2 *;* (c) $\eta_2^{(i)}$ *;* (d) η_3 *($NV \simeq 427$).*

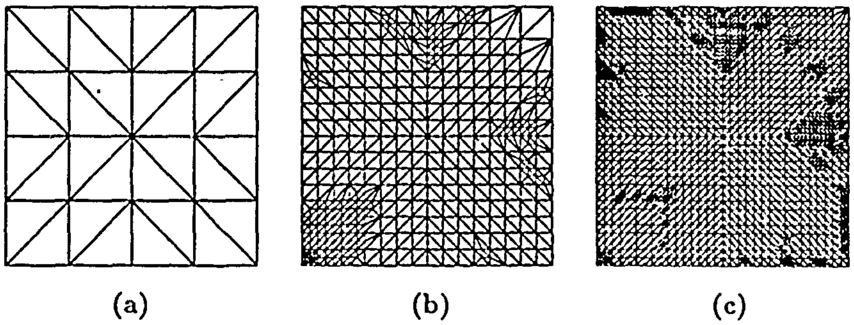


FIG. 5. *A sequence of refined grids on a smooth problem leads to quasi-uniform meshes* (a) *initial grid $NV = 25$;* (b) *$NV = 394$;* (c) *$NV = 1496$.*

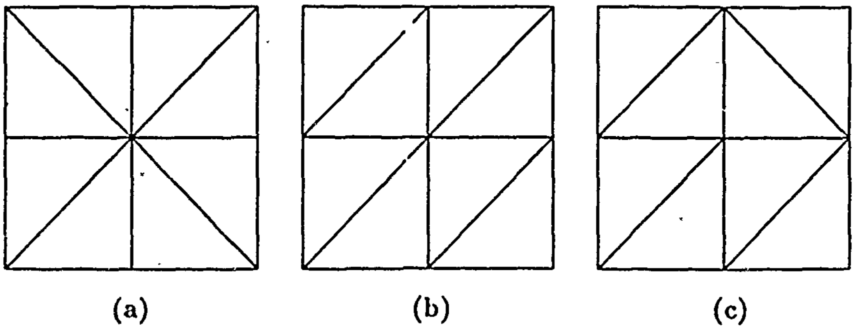


FIG. 6. *Different initial grids for the driven cavity problem.*

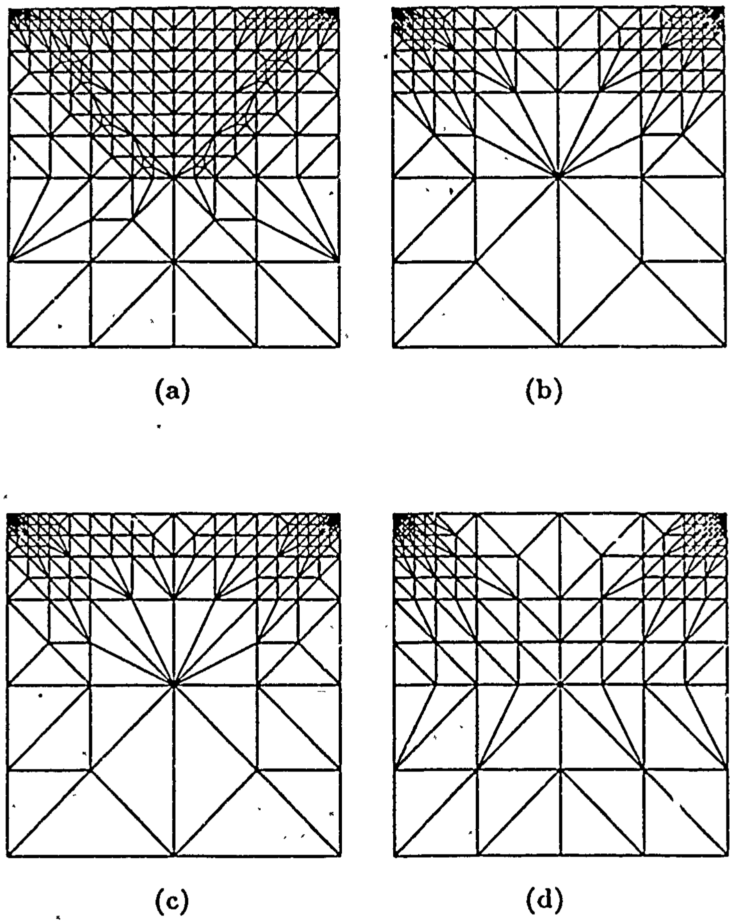


FIG. 7. *Adapted grids using estimates* (a) η_1 ; (b) η_2 ; (c) $\eta_2^{(i)}$; (d) η_3 *(*$NV \simeq 250$*).*

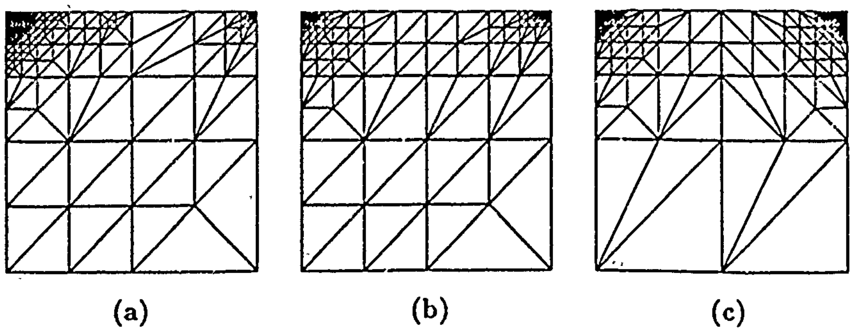
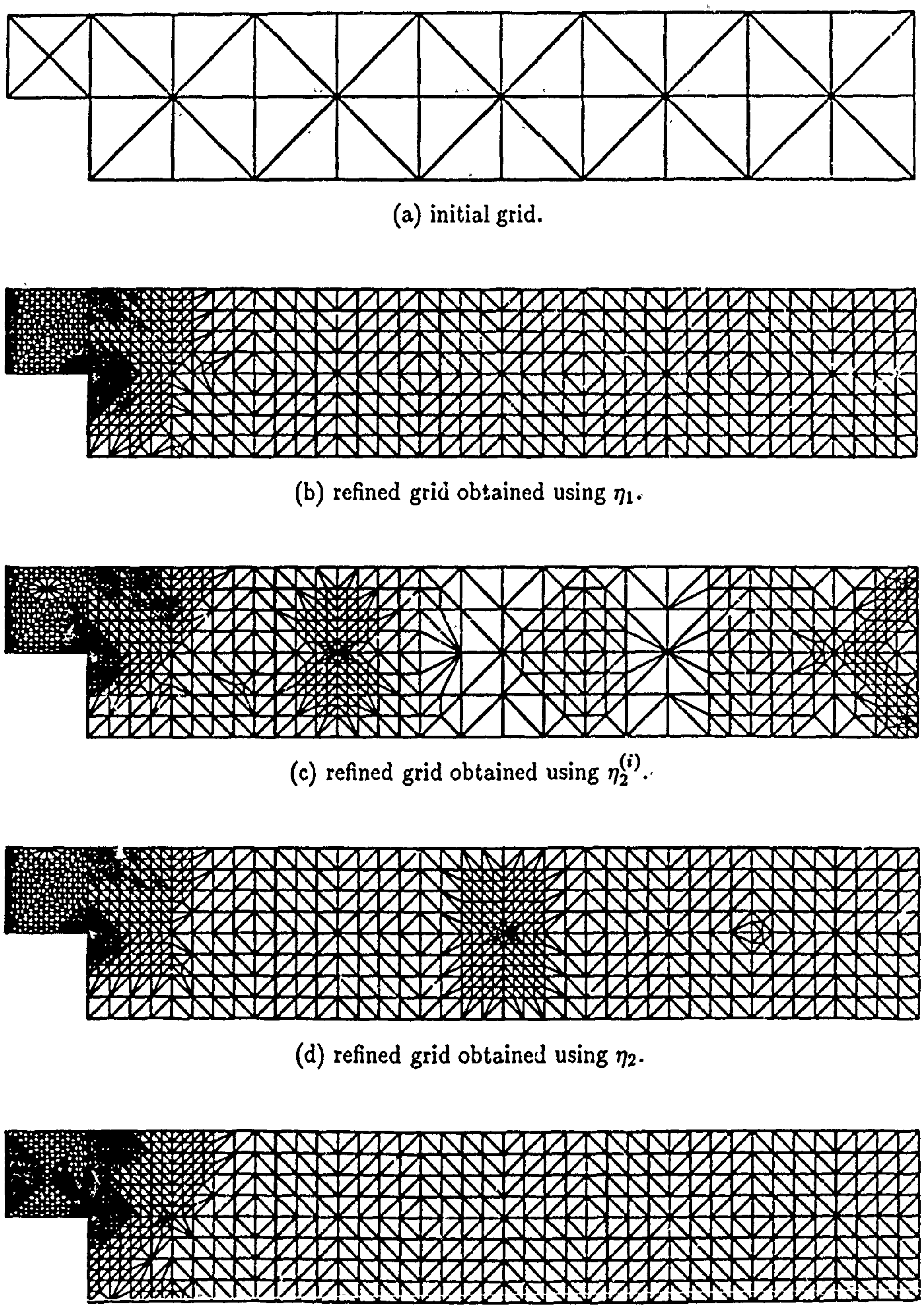


FIG. 8. (a) *grid obtained by refinement of the initial grid 6*(b); (b) *same as* (a), *but the estimate is now computed from a solution calculated more often during the refinement process;* (c) *refined grid starting from the coarse mesh 6*(c).



(a) initial grid.

(b) refined grid obtained using η_1.

(c) refined grid obtained using $\eta_2^{(i)}$.

(d) refined grid obtained using η_2.

(e) refined grid obtained using η_3.

FIG. 9.

PREDICTOR-CORRECTOR PROCEDURES FOR STRESS AND FREE VIBRATION ANALYSES OF MULTILAYERED COMPOSITE PLATES AND SHELLS

Ahmed K. Noor, W. Scott Burton and Jeanne M. Peters
George Washington University
NASA Langley Research Center
Hampton, Virginia 23665

To be published in Computer Methods in Applied
Mechanics and Engineering

January 1990

PREDICTOR-CORRECTOR PROCEDURES FOR STRESS AND FREE VIBRATION ANALYSES OF MULTILAYERED COMPOSITE PLATES AND SHELLS

Ahmed K. Noor, W. Scott Burton and Jeanne M. Peters
George Washington University
NASA Langley Research Center
Hampton, Virginia 23665

ABSTRACT

A study is made of two predictor-corrector procedures for the accurate determination of the global, as well as detailed, static and vibrational response characteristics of plates and shells. Both procedures use first-order shear deformation theory in the predictor phase, but differ in the elements of the computational model being adjusted in the corrector phase. The first procedure calculates *a posteriori* estimates of the composite correction factors and uses them to adjust the transverse shear stiffnesses of the plate (or shell). The second procedure calculates *a posteriori* the functional dependence of the displacement components on the thickness coordinate. The corrected quantities are then used in conjunction with the three-dimensional equations to obtain better estimates for the different response quantities. Extensive numerical results are presented showing the effects of variation in the geometric and lamination parameters for antisymmetrically laminated anisotropic plates, and simply supported multilayered orthotropic cylinders, on the accuracy of the linear static and free vibrational responses obtained by the predictor-corrector procedures. Comparison is also made with the solutions obtained by other computational models based on two-dimensional shear deformation theories. For each problem the standard of comparison is taken to be the analytic three-dimensional elasticity solution. The numerical examples clearly demonstrate the accuracy and effectiveness of the predictor-corrector procedures.

NOMENCLATURE

$\bar{a}_{31}, \bar{a}_{32}, \bar{a}_{33}$	three-dimensional elastic compliance coefficients
$\bar{a}_{36}, \bar{a}_{44}, \bar{a}_{45}, \bar{a}_{55}$	
$c_{11}, c_{12}, c_{22}, c_{16}, c_{26}, c_{66}$	plane stress reduced stiffness coefficients of different layers

$c_1, c_2, c_3, c_w, c_{u1}, c_{u2}$	constants of integration (see Eqs. 2 to 4 and 7 to 9)
E_L, E_T	elastic moduli in direction of fibers and normal to it
f_1, f_2, f_3	body force components in coordinate directions
G_{LT}, G_{TT}	shear moduli in plane of fibers and normal to it
h	total thickness of plate (or shell)
k_1^0, k_2^0, k_1, k_2	initial and corrected values of composite correction factors
L_1, L_2	side lengths in the x_1 and x_2 directions (for the cylinder $L_2 = 2\pi r_0$)
$M_{\alpha\beta}$ $(\alpha,\beta = 1,2)$	bending stress resultants
m,n	Fourier harmonics in the x_1 and x_2 directions
NL	total number of layers in the plate (or shell)
$N_{\alpha\beta}$ $(\alpha,\beta = 1,2)$	extensional stress resultants
p_1, p_2, p	intensities of external forces in coordinate directions
p_0	intensity of normal loading on plate (or cylinder)
Q_α $(\alpha=1,2)$	transverse shear stress resultants
r	parameter which equals 1 and λr_0 for plates and cylinders, respectively (see Eq. 9)
r_0	radius of middle surface of cylinder
U	total strain energy of plate (or shell)
U_1	strain energy component associated with $\sigma_{11}, \sigma_{22}, \sigma_{12}$ and $\varepsilon_{11}, \varepsilon_{22}, 2\varepsilon_{12}$
U_2	strain energy component associated with σ_{13}, σ_{23} and $2\varepsilon_{13}, 2\varepsilon_{23}$
U_3	strain energy component associated with σ_{33} and ε_{33}
u_α, w $(\alpha=1,2)$	displacement components in the x_α, x_3 coordinate directions
x_α, x_3 $(\alpha=1,2)$	orthogonal coordinate system for the plate (or shell)
$\left.\begin{matrix}\varepsilon_{11}^0, \varepsilon_{22}^0, 2\varepsilon_{12}^0 \\ \kappa_{11}^0, \kappa_{22}^0, 2\kappa_{12}^0\end{matrix}\right\}$	extensional and bending strain components of the middle surface of the plate (or shell)

$2\varepsilon_{13}$, $2\varepsilon_{23}$	transverse shear strain components
ε_{33}	transverse normal strain component
$\zeta = x_3/h$	dimensionless transverse coordinate
κ_0	initial curvature of the middle surface of the cylinder (for plates κ_0 =0)
λ	$= 1 + x_3/r_0$ for cylinders = 1 for plates
ν_{LT}, ν_{TT}	Poisson's ratio for the material of individual layers
$\xi_\alpha = x_\alpha/L_\alpha$	(α=1,2 and is not summed) dimensionless surface coordinates
ρ	mass density of material
σ , σ_{22}, σ_{33}	normal stress components
σ_{12}, σ_{13}, σ_{23}	shear stress components
ϕ_α (α=1,2)	rotation components
ω	frequency of vibration of the plate (or shell)
∂_α	$= \partial/\partial x_\alpha$

Superscripts

k = layer number

o = predictions of first-order shear deformation theory

— =(bar over a symbol) response quantities obtained by using three-dimensional elasticity theory

Subscripts

E = three-dimensional elasticity solution

L = direction of fibers

m,n = Fourier harmonics in x_1 and x_2 coordinate directions

T = direction normal to fibers

α,β=1,2

1. INTRODUCTION

Since the publication of the first monographs on anisotropic plates and shells in 1947 and 1961 (Refs. 1 and 2), considerable progress has been made in the analysis of laminated and anisotropic plates and shells. Review of the many contributions on this subject is given in a number of monographs (see, for example, Refs. 3 through 17) and survey papers (Refs. 18 to 24).

Most of the early publications were limited to predicting gross response characteristics (vibration frequencies, buckling loads, average through-the-thickness displacements and rotations) of thin laminated plates and shells. The classical lamination theory, based on neglecting transverse shear strains and transverse normal strains in the plate (or shell) (see, for example, Refs. 3, 13 and 25) is adequate for this purpose. The expanded use of fibrous composite materials in high-technology industries (aircraft, automotive, shipbuilding and other industries) has stimulated interest in the accurate prediction of the detailed response and failure characteristics of laminated anisotropic plates and shells. Several modeling approaches have been proposed which take into account the relatively low elastic moduli in the lateral and transverse directions. Some of these modeling approaches are extensions of similar approaches used for isotropic plates and shells and include: 1) three-dimensional and quasi-three-dimensional elasticity models (Refs. 26 to 33); 2) first-order shear deformation theories based on linear displacement and/or piecewise linear stress variation through-the-thickness of the entire laminate (see, for example, Refs. 34 to 38); and 3) higher-order shear-deformation theories based on nonlinear (or piecewise linear) variation of displacements; and/or nonlinear variation of stresses through the thickness (Refs. 14 and 39 to 47).

In quasi-three-dimensional models simplifying assumptions are made regarding the stress (or strain) state in the laminate (or in the individual layers), but no *a priori* assumptions are made about the distribution of the different response characteristics in the thickness direction. The use of both three-dimensional and quasi-three-dimensional models for predicting the response characteristics of laminated anisotropic plates and shells with complicated geometry is computationally expensive, and therefore, may not be feasible for practical composite plates and shells.

Experience with two-dimensional shear deformation theories has shown them to be inadequate for the accurate prediction of transverse stresses and deformations. This is particularly true when first-order theories (in which the transverse shear strains are assumed to be constant in the thickness direction (Ref. 4)) are used for analyzing medium-thick and thick plates and shells. The range of validity of first-order shear deformation theory was found to be strongly dependent on the factors used in adjusting the transverse shear stiffnesses of the laminate (see Refs. 22, 48 and 49).

A simple approach for the accurate evaluation of transverse stresses and strains in medium-thick composite plates and shells is to use a two-dimensional shear deformation theory for calculating the in-plane stresses, and then the three-dimensional equilibrium equations to determine the transverse stresses. An improvement of this approach was proposed (Refs. 33 and 49) in which better estimates are obtained for the transverse shear stiffnesses and then used to correct the gross response characteristics, which are in turn used in evaluating the transverse stresses. The present study extends *the idea of using the information obtained from a simple two-dimensional shear deformation theory to correct certain key elements of the computational model (in an inexpensive, postprocessing mode), and hence, improve the response predictions.* Specifically, the objectives of the present paper are: a) to assess the accuracy of two predictor-corrector procedures for calculating the detailed response characteristics of multilayered composite plates and shells, and b) to discuss the potential of using these procedures in solving practical plate and shell problems.

Both predictor-corrector procedures use first-order shear deformation theory in the predictor phase but differ in the elements of the computational model being corrected, namely: a) correcting transverse shear stiffnesses; or b) correcting the thickness distribution of displacements. The first procedure is described in Refs. 33 and 49, and the details of the second procedure are presented here for the first time.

The composite plates and shells considered herein consist of a number of perfectly bonded layers. The individual layers are assumed to be homogeneous. At each point of the structure a plane of elastic symmetry exists, parallel to the middle surface. The sign convention for the different displacement and stress components is shown in Fig. 1. Both antisymmetrically laminated composite plates and orthotropic multilayered cylinders are considered.

In order to obtain analytic solutions for the antisymmetrically laminated plates, each of the displacement parameters, strain components and stress resultants is decomposed into symmetric and antisymmetric components in the thickness coordinate. Different double Fourier series expansions, in the surface coordinates, are used for the symmetric and antisymmetric components (see Refs. 50 and 51). For both the plates and cylinders the solutions and the external surface loads are assumed to be periodic in the surface coordinates. Extensive numerical results are presented showing the effects of variation in the lamination and geometric parameters of the pl'tes and cylinders on the accuracy of the static and vibrational responses predicted by the first-order shear-deformation theory, as well as by the two predictor-corrector approaches. The standard of comparison is taken to be the exact three-dimensional elasticity solutions.

2. BASIC IDEA OF PREDICTOR-CORRECTOR PROCEDURES

The predictor-corrector procedures used in the present study are iterational processes in which the information obtained in the first (predictor) phase of the analysis is used to correct key elements of the computational model, and hence, improve the response predictions. Numerical experiments have shown that only one iteration is needed (in the corrector phase) to obtain highly accurate response predictions.

Two predictor-corrector procedures are considered in the present study, both use first-order shear-deformation theory in the predictor phase to calculate initial estimates for the gross response characteristics of the structure (vibration frequencies, average through-the-thickness displacements and rotations), as well as the in-plane stresses; then three-dimensional equilibrium equations and constitutive relations are used to calculate transverse shear and transverse normal stresses and strains.

The two procedures differ in the elements of the computational model being adjusted in the corrector phase. The first procedure calculates *a posteriori* estimates of the composite correction factors and uses them to adjust the transverse shear stiffnesses of the plate (or shell). By contrast, the second procedure calculates *a posteriori* the functional dependence of the displacement components on the thickness coordinate. The corrected quantities are then used in conjunction with the three-dimensional equations to obtain better estimates for the different response quantities. A schematic representation of the different steps involved in the two predictor-corrector procedures is given in Fig. 2. The details of the first procedure are given in Refs. 33 and 49, and the application of the second procedure to static and free vibration problems of multilayered composite plates and cylindrical shells is outlined subsequently. The superscript o refers to the predictions of the first-order shear deformation theory; and a bar (-) over a symbol refers to the response quantities obtained by using three-dimensional elasticity equations.

2.1 Static Analysis

The sequence of steps involved in the calculation of stresses and displacements are as follows:

A. Predictor Phase

A first-order shear deformation theory is used with an initial set of composite correction factors k^{o}_{α} to evaluate through-the-thickness displacements u^{o}_{α}, w^{o}; rotation components ϕ^{o}_{α}; middle surface strains and curvature changes $\varepsilon^{o}_{\alpha\beta}$, $\kappa^{o}_{\alpha\beta}$; average transverse shear strains $2\varepsilon^{o}_{\alpha 3}$; and stress resultants $N^{o}_{\alpha\beta}$, $M^{o}_{\alpha\beta}$, Q^{o}_{α} ($\alpha,\beta = 1,2$). The sign convention for the generalized displacements and stress resultants is shown in Fig. 1.

Then the in-plane stresses in the kth layer, $\sigma^{o}_{\alpha\beta}$, are calculated by using the equations:

$$\begin{Bmatrix} \sigma^{o}_{11} \\ \sigma^{o}_{22} \\ \sigma^{o}_{12} \end{Bmatrix}^{(k)} = \begin{bmatrix} c_{11} & c_{12} & c_{16} \\ & c_{22} & c_{26} \\ \text{Symm} & & c_{66} \end{bmatrix}^{(k)} \left(\begin{Bmatrix} \varepsilon^{o}_{11} \\ \varepsilon^{o}_{22} \\ 2\varepsilon^{o}_{12} \end{Bmatrix} + x_3 \begin{Bmatrix} \kappa^{o}_{11} \\ \kappa^{o}_{22} \\ 2\kappa^{o}_{12} \end{Bmatrix} \right) \tag{1}$$

in which the c's are the plane stress reduced stiffness coefficients of the kth layer (see Refs. 52 and 53). For convenience the superscript (k) will be dropped in most of the succeeding steps.

Next, the transverse shear and normal stresses are obtained by integrating the three-dimensional equilibrium equations in the thickness direction as follows:

$$\bar{\sigma}_{13} = -\frac{1}{\lambda} \int_{-h/2}^{x_3} \left[\lambda \partial_1 \sigma^{o}_{11} + \partial_2 \sigma^{o}_{12} + \lambda f_1 \right] dx_3 + c_1 \tag{2}$$

$$\bar{\sigma}_{23} = -\frac{1}{\lambda^2} \int_{-h/2}^{x_3} \lambda \left[\lambda \partial_1 \sigma^{o}_{12} + \partial_2 \sigma^{o}_{22} + \lambda f_2 \right] dx_3 + c_2 \tag{3}$$

$$\bar{\sigma}_{33} = -\frac{1}{\lambda}\int_{-h/2}^{x_3}\left[\lambda\,\partial_1\bar{\sigma}_{13} + \partial_2\bar{\sigma}_{23} - \kappa_o\,\sigma_2^o + \lambda f_3\right]dx_3 + c_3 \tag{4}$$

where $\lambda=1$ for plates, and $\lambda = 1+ x_3/r_o$ for cylinders; $\kappa_o=0$ for plates; and $\kappa_o = 1/r_o$ for cylinders (r_o is the radius of the middle surface); c_1 , c_2 and c_3 are integration constants obtained from the stress conditions at the outer surfaces of the laminate; and f_1 , f_2 and f_3 are body force components in the coordinate directions; and $\partial_\alpha = \partial/\partial x_\alpha$. Note that because of the discontinuity of σ_{11}^o , σ_{22}^o and σ_{12}^o at layer interfaces, the integrations in Eqs. 2 to 4 are performed in a piecewise manner (layer by layer).

The transverse shear and transverse normal strains are obtained from the following three-dimensional constitutive relations:

$$\begin{Bmatrix}2\bar{\varepsilon}_{13}\\2\bar{\varepsilon}_{23}\end{Bmatrix} = \begin{bmatrix}\bar{a}_{55} & \bar{a}_{45}\\ \bar{a}_{45} & \bar{a}_{44}\end{bmatrix}\begin{Bmatrix}\bar{\sigma}_{13}\\ \bar{\sigma}_{23}\end{Bmatrix} \tag{5}$$

$$\bar{\varepsilon}_{33} = \bar{a}_{31}\,\sigma_{11}^o + \bar{a}_{32}\,\sigma_{22}^o + \bar{a}_{36}\,\sigma_{12}^o + \bar{a}_{33}\,\bar{\sigma}_{33} \tag{6}$$

in which the a' s are three-dimensional compliance coefficients (see Refs. 52 and 53).

The distribution of the displacement components in the thickness direction is obtained by integrating the three-dimensional transverse strain-displacement relationships as follows:

$$\bar{w} = -\frac{1}{\lambda}\int_{-h/2}^{x_3}\bar{\varepsilon}_{33}\,dx_3 + c_w \tag{7}$$

$$\bar{u}_1 = \int_{-h/2}^{x_3}(2\bar{\varepsilon}_{13} - \partial_1\bar{w})\,dx_3 + c_{u1} \tag{8}$$

$$\frac{\bar{u}_2}{r} = \int_{-h/2}^{x_3}\frac{1}{r}(2\bar{\varepsilon}_{23} - \partial_2\bar{w})\,dx_3 + c_{u2} \tag{9}$$

where $r=1$ for plates, and $r=\lambda r_o$ for cylinders; and c_w, c_{u1} ; c_{u2} and are integration constants, which are obtained by setting $\bar{u}_1$, $\bar{u}_2$ and $\bar{w}$ at $x_3=0$ to be equal to the corresponding displacements of the first-order shear deformation theory.

B. Corrector Phase

The calculation of the corrected response characteristics of the plate (or shell) may be conveniently divided into three steps, namely: 1) generation of coordinate (basis) displacement functions; 2) computation of amplitudes of the coordinate functions, and evaluation of corrected through-the-thickness distribution of displacements; and 3) calculation of through-the-thickness distributions of in-plane strains, in-plane stresses, and transverse stresses using the three-dimensional strain-displacement, constitutive relations, and equilibrium equations. The first two steps are described subsequently.

1. *Generation of Coordinate (Basis) Displacement Functions.* Each of the three displacement components $\bar{u}_1, \bar{u}_2, \bar{w}$, Eqs. 7 to 9, is decomposed into symmetric and antisymmetric functions of the thickness coordinate x_3; each of the symmetric components is further subdivided into a constant component (value of the displacement at the middle surface) and a nonlinear function in x_3. Similarly, each of the antisymmetric components is decomposed into a linear and a nonlinear function of x_3. In the present study the linear functions for the in-plane displacements were chosen to be the average through-the-thickness rotation components (used in the first-order shear deformation theory), and the linear function for the transverse displacement was chosen such that the value of the nonlinear component at the outer surfaces was zero.

2. *Computation of Amplitudes of the Coordinate Functions.* The resulting four symmetric/antisymmetric functions, associated with each of the displacement components, $\bar{u}_1, \bar{u}_2, \bar{w}$, are now chosen as coordinate (or basis functions), and the displacement component is expressed as a linear combination of the four functions, with unknown parameters (representing the amplitudes of the coordinate functions, or displacement modes).

The twelve unknown parameters are obtained by using the Rayleigh-Ritz technique in conjunction with the minimum potential energy principle (based on the three-dimensional elasticity equations).

2.2 Free Vibration Analysis

In the predictor phase the first-order shear deformation is used to obtain estimates for the eigenvalues $(\omega^{\circ})^2$, average through-the-thickness modal displacements, and rotation components as well as the in-plane stresses. The body force components f_1, f_2 and f_3 in Eqs. 2 to 4 are set equal to the inertia forces, i.e.,

$$\begin{Bmatrix} f_1 \\ f_2 \\ f_3 \end{Bmatrix} = \rho\left(\omega^{o}\right)^2 \begin{Bmatrix} u_1^o + x_3\ \phi_1^o \\ u_2^o + x_3\ \phi_2^o \\ w^o \end{Bmatrix} \tag{10}$$

in which ρ is the mass density of the material. Equations 5 to 9 are then used to obtain the distributions of transverse strains and displacements in the thickness direction.

In the second step of the corrector phase, Hamilton's principle is used to generate a generalized matrix eigenvalue problem from which the corrected vibration frequency and eigenvectors (vectors of amplitudes of displacement coordinate functions) are calculated.

2.3 Comments on the Predictor-Corrector Computational Procedure

The following comments on the computational procedure are in order:

1. Because of the assumed through-the-thickness linear distribution of strains in the first-order shear deformation theory, and the associated piecewise linear distribution of stresses, the transverse stresses obtained from Eqs. 2 to 4 *may not satisfy all the stress conditions at the top and bottom surfaces, and at layer interfaces.* This is particularly true for laminated cylinders. The accuracy of the transverse stresses obtained by the predictor-corrector procedure may be somewhat sensitive to which conditions are satisfied. Numerical experiments have shown that good accuracy is obtained when the stress conditions at both the top and bottom surfaces are satisfied, and the discontinuities in the transverse stresses occur at or near the middle surface. These transverse stress discontinuities can be eliminated by using an error distribution procedure. Such a procedure was not used in the present study.

2. In the corrector phase a mixed formulation can be used in which the fundamental unknowns consist of the three displacement components and the three transverse stress components. The appropriate variational principle to be used in the computation of amplitudes of coordinate functions is that proposed by Reissner (Refs. 54 and 55).

3. NUMERICAL STUDIES

To assess the accuracy and effectiveness of predictor-corrector computational procedures, a large number of stress and free vibration problems of multilayered composite plates and cylinders have been solved by these techniques. The composite plates considered in the present study are square laminates with $L_1=L_2=1.0$, and have antisymmetric lamination with respect to the middle plane. Both the static loading, and the solutions considered, are periodic in x_1 and x_2 with periods $2L_1$ and $2L_2$ for plates ($2L_1$ and $2L_2$ for cylinders). The composite cylinders considered are simply supported laminated circular cylinders. The fibers of the different layers alternate between the circumferential and longitudinal directions, with the fibers of the top layers running in the circumferential direction.

The material characteristics of the individual layers were taken to be those typical of high-modulus fibrous composites, namely:

$$E_L/E_T=15,\ G_{LT}/E_T=0.5,\ G_{TT}/E_T=0.3356,\ \nu_{LT}=0.3,\ \nu_{TT}=0.49$$

where subscript L refers to the direction of fibers and subscript T refers to the transverse direction; and ν_{LT} is the major Poisson's ratio. For static stress analysis problems the plates and cylinders were subjected to sinusoidal normal loading. For plates the loading was antisymmetric in the x_3 direction and was normal to the top and bottom surfaces. The total normal loading $p=p_o \sin \pi\xi_1 \sin \pi\xi_2$. For cylinders the loading was normal to the inner surface of the cylinder and had the form $p=p_o \sin \pi\xi_1 \cos 2\pi n\xi_2$. For free vibration problems only the lowest frequencies for each pair of m,n were considered along with the associated mode shapes and modal stresses.

For each problem, the solutions obtained by the predictor-corrector procedures were compared with the predictions of five different modeling approaches based on two-dimensional shear-deformation plate and shell theories. The five modeling approaches are: first-order shear deformation theory based on linear variation of u_α and constant w through the thickness; first-order theory based on linear variation of u_α and w through the thickness; higher-order theory with quintic variation of u_α and w through the thickness; simplified higher-order theory based on cubic variation of u_α and constant w through the thickness, with the conditions of zero transverse shear stresses imposed at the outer surfaces of the plate (or shell) to reduce the number of generalized displacement parameters; and discrete-layer theory with piecewise linear distribution of u_α and constant w in the thickness direction. The modeling approaches considered are listed in Table 1, and will henceforth be referred to as models 1 through 5. The predictor-corrector procedures will be referred to as models 6 and 6A (see Table 1). The standard of comparison is taken to be the analytic three-dimensional elasticity solutions. The methods of obtaining these solutions are outlined in Refs. 39, 50 and 56 for plates and cylinders, respectively.

For plates, three parameters were varied, namely, the thickness ratio of the plate, h/L_1; the number of layers, NL; and the fiber orientation angle of the individual layers, θ. The thickness ratio was varied between 0.01 and 0.4; the number of layers was varied between 2 and 20: and θ was varied between 0° and 45°. The wave numbers in the x_1 and x_2 directions were selected to be 1, and the aspect ratio, L_1/L_2, was also selected to be 1.0. For cylinders, three parameters were varied, namely, the number of layers, NL; the thickness ratio h/r_0; and the circumferential wave number, n. The longitudinal wave number was selected to be 1, and the length-to-radius ratio, L/r_0, was also selected to be 1.0. The number of layers was varied between 2 and 20; h/r_0 between 0.01 and 0.3, and n between 0 and 10.

As a step towards establishing the range of validity of the predictor-corrector procedures, the total strain energy of the structure was decomposed into three components: U_1 associated with $\sigma_{\alpha\beta}$ and $\varepsilon_{\alpha\beta}$; U_2 associated with $\sigma_{\alpha 3}$ and $2\varepsilon_{\alpha 3}$; and U_3 associated with σ_{33} and ε_{33} $\left(U_3 = \frac{1}{2}\int \sigma_{33}\,\varepsilon_{33}\,dV\right)$. The total strain of the structure $U = U_1 + U_2 + U_3$. The assessment of the accuracy of the predictor-corrector procedures included both global response characteristics (vibration frequencies and strain energy components), as well as detailed stress and displacement distributions in the thickness direction.

Typical results are presented in Figs. 3 to 5, Tables 2 and 3 for the antisymmetrically laminated plates, and in Figs. 6 to 8 and Tables 4 and 5 for the simply-supported cylinders.

The effects of variation of the three parameters h/L_1, NL and θ for plates (and NL, h/r_0 and n for cylinders) on the minimum vibration frequencies, and the energy components U_1, U_2, U_3 obtained by the three-dimensional elasticity model are depicted in Table 2 and Fig. 3 for plates and in Table 4 and Fig. 6 for cylinders.

An indication of the accuracy of the minimum vibration frequencies and energy components obtained by the predictor-corrector procedures, and other computational models listed in Table 1, is given in Figures 4, 5, 7 and 8, and in Tables 2 and 4. Figures 4 and 7 show the effects of h/L_1 for plates and, h/r_0 and n for cylinders, on the accuracy of the strain energy components obtained by different models. Tables 2 and 4 give an indication of the effect of different plate and cylinder parameters on the accuracy of the minimum frequencies obtained by the different models. Figures 5 and 8 give an indication of the accuracy of the displacement, stress, and transverse shear strain energy distributions predicted by different models. In Figure 5 both the symmetric and antisymmetric parts of the response quantities (with respect to the middle plane) are shown. The symmetric $u_{1,S}$, $\sigma_{11,S}$, $\sigma_{33,S}$, and antisymmetric $\sigma_{13,A}$, $U_{13,A}$ are shaded in Fig. 5. The antisymmetric w_A are approximately zero. Note that since the symmetric and antisymmetric components of each response quantity are multiplied by different trigonometric functions in x_1 and x_2 (see Refs. 50 and 51), the value of the response quantity is a linear combination of the two components. In Tables 3 and 5 the maximum absolute values are given for the displacements, stresses and transverse shear strain energy densities obtained by the three-dimensional elasticity model for plates and cylinders.

An examination of Tables 2 and 4 and Figures 3 to 8 reveals:

1. The transverse shear strain energy ratio, U_2/U, increases with the increase in h/L_1, θ and NL for plates (h/r_0, n and NL for cylinders). The increase in U_2/U is associated with a decrease in the ratio U_1/U. On the other hand, for all the vibration problems considered, U_3/U was found to be very small (less than 1%). For statically loaded plates, U_3/U approaches 2.2% for thick multilayered plates with $h/L_1=0.3$ and $NL\geq10$ (Fig. 3) and 36.9% for multilayered cylinders with $h/r_0=0.3$, $n\geq8$ and $NL\geq10$ (see Fig. 6).

2. As expected, the accuracy of the first-order shear deformation theory (models 1 and 1A) decreases as h/L_1, and θ increase for plates (h/r_0 and n for cylinders). The range of validity of the first-order theory is strongly dependent on the values of the composite shear correction factors used, k_1^0 and k_2^0 (see Tables 2 and 4). The errors in the predictions of model 1A (with k_1 and k_2 computed from the cylindrical bending condition of Refs. 57 and 58) are considerably

lower than those of model 1.

3. Despite the larger number of displacement parameters of model 2, its predictions are generally less accurate than those of model 1 (see Figures 4 and 7a). This is attributed to the assumption of constant transverse normal strain, and piecewise constant transverse normal stresses, through the thickness in model 2, which result in considerably overestimating the in-plane stresses $\sigma_{\alpha\beta}$. An exception to that is the case of statically loaded thick cylinders with $h/r_o \geq 0.2$. Because of the importance of transverse normal stresses, the predictions of model 2 are more accurate than those of model 1 (see Fig. 7b).

4. For the entire range of parameters considered, the global response characteristics predicted by the higher-order shear deformation theory (model 3) are fairly accurate (see Tables 2 and 4 and Figures 4 and 7). However, the distribution of transverse stresses through the thickness obtained by model 3 is not as accurate as the gross response characteristics (results not shown).

5. For plates with $h/L_1 \leq 0.2$ (and cylinders with $h/r_o \leq 0.2$ and $n \leq 8$) the gross response characteristics predicted by the simplified higher-order theory (model 4) are fairly accurate. A rapid degradation in accuracy as the ratio h/L_1 for plates (or the ratio h/r_o for cylinders) increases beyond 0.2.

6. When the transverse normal stresses σ_{33} are not significant (or the ratio U_3/U is small), the global response characteristics predicted by the discrete-layer theory (model 5) are fairly accurate (see Figs. 4 and 7). Note that for $NL \geq 8$ the number of displacement parameters used in this model exceed those used in all other models.

7. The predictor-corrector procedures (models 6 and 6A) appear to be very effective approaches for the accurate determination of the global, as well as detailed response characteristics of plates and cylinders. This is particularly true for model 6A which generates very accurate transverse stresses, even for very thick shells (see Figs. 5 and 8). Specifically, the following three observations on model 6 can be noted:

a) The numerical values of the corrected composite shear correction factors, k_1 and k_2, are fairly insensitive to their initial values, k_1^o and k_2^o, used in the first-order shear deformation theory. They depend on the distributions of the transverse shear strains in the thickness direction which, in turn, are functions of both the lamination and geometric parameters of the plate (or cylinder).

b) If k_1^0 and k_2^0 are both selected to be 1, the error in the global response quantities obtained in the first (predictor) phase, for plates with $h/L_1 \geq 0.2$ (or cylinders with $h/r_0 \geq 0.2$ and $n \geq 4$), may be unacceptable; however, the corrector phase improves these predictions substantially, and results in highly accurate distributions of displacements and stresses through the thickness (see Figs. 4, 5, 7 and 8).

c) The accuracy of the response quantities obtained using the predictor-corrector procedure is insensitive to the initial shear correction factors selected. It is also insensitive to the selection of the reanalysis procedure in the correction phase. For example, when the calculated composite correction factors are much different from their initial values, a first-order Taylor series approximation (with respect to the composite correction factors) provides sufficiently accurate estimates for the response quantities (see Ref. 51).

The aforementioned observations point to the fact that accurate prediction of the distribution of stresses and displacements through the thickness of multilayered plates and cylinders .quires the use of three-dimensional equilibrium and constitutive relations. These equations can be used in an inexpensive, postprocessing mode with any of the modeling approaches based on two-dimensional theories. The predictor-corrector procedures have the advantage of starting with a simple first-order theory in the first phase to obtain estimates for the global response characteristics, and then modifying the key elements of the computational model before calculating the displacement distribution in the thickness direction.

4. POTENTIAL OF THE PREDICTOR-CORRECTOR PROCEDURES

The predictor-corrector procedures appear to have high potential for the accurate prediction of vibration frequencies, stresses and deformations in multilayered composite plates and shells. The numerical studies conducted for antisymmetrically laminated anisotropic plates and simply supported orthotropic cylinders demonstrated the accuracy and effectiveness of the predictor-corrector procedures. In particular, the following points are worth mentioning:

1. The predictor-corrector procedures can be applied, in conjunction with finite element models, to the analysis of anisotropic plates and shells with arbitrary geometry. The calculation of the transverse stresses, and the correction phase (including the calculation of composite shear correction factors in model 6 and the thickness distribution of the in-plane and transverse displacements in model 6A) can be performed on the element level for selected elements (in the critical regions of the plate and shell models).

2. Although any of the two-dimensional shear-deformation plate and sheil theories can be used in the first (predictor) phase of the predictor-corrector procedures, the first-order shear deformation theory has the following two major advantages over other theories: a) only five displacement parameters are used to describe the deformation; and b) in the finite element implementation only C^0 continuity is required. The simplified higher-order theories (model 4) and a simplified discrete-layer theory (see Refs. 22 and 24) share the first advantage, but require C^1 continuity in their finite element implementation.

5. CONCLUDING REMARKS

A study is made of two predictor-corrector procedures for the accurate determination of the global as well as detailed response characteristics of plates and shells. Both procedures can be thought of as iterational processes in which the information obtained in the first (predictor) phase of the analysis is used to correct key elements of the computational model, and hence, improve the response predictions. The two predictor-corrector procedures use first-order shear-deformation theory in the predictor phase to calculate initial estimates for the gross response characteristics of the structure (vibration frequencies, average through-the-thickness displacements and rotations), as well as in-plane stresses; then three-dimensional equilibrium equations and constitutive relations are used to calculate transverse shear and transverse normal stresses and strains. The two procedures differ in the elements of the computational model being adjusted in the corrector phase. The first procedure calculates *a posteriori* estimates of the composite correction factors and uses them to adjust the transverse shear stiffnesses of the plate (or shell). The second procedure calculates *a posteriori* the functional dependence of the displacement components on the thickness coordinate. The corrected quantities are then used in conjunction with the three-dimensional equations to obtain better estimates for the different response quantities.

Extensive numerical results are presented for multilayered antisymmetrically laminated plates and simply-supported orthotropic cylinders, showing the effects of variation in the geometric and lamination parameters on the accuracy of the static and free vibrational responses obtained by the predictor-corrector procedures. For each problem the standard of comparison is taken to be the analytic three-dimensional elasticity solution. Comparison is also made with solutions obtained by five computational models based on first-order as well as higher-order two-dimensional shear deformation theories. The five computational models are: first-order shear deformation theory based on linear variation of u_α and constant w through the thickness; first-order theory based on linear variation of u_α and w through the thickness; higher-order theory with quintic variation of u_α and w through the thickness; simplified higher-order theory based on cubic variation of u_α and constant w through the thickness, with the conditions of zero transverse shear stresses imposed at the outer surfaces of the plate (or shell) to reduce the number of generalized displacement parameters; discrete-layer theory with piecewise linear distribution of u_α and constant w in the thickness direction.

On the basis of the numerical results the following conclusions are justified:

1. As has been reported previously (Refs. 22, 24 and 48), the accuracy of the predictions of first-order shear deformation theory is strongly dependent on the values of the composite-correction factors. The use of the composite shear correction factors proposed in Refs. 57 and 58 results in fairly accurate gross response characteristics for a wide range of lamination and geometric parameters.

2. The gross response characteristics predicted by higher-order shear deformation theories are fairly accurate for a wide range of geometric and lamination parameters. However, the accurate prediction of the transverse stress and displacement distributions through the thickness in multilayered plates and shells requires the use of three-dimensional equilibrium and constitutive relations. These equations can be used in an inexpensive, postprocessing mode with any of the modeling approaches based on two-dimensional theories.

3. The predictor-corrector procedures appear to be very effective procedures for the

accurate determination of the global as well as the detailed response characteristics of plates and shells. The accuracy of the response quantities obtained in the first (predictor) phase of laminates with thickness-to-wavelength (of the deformation) ratio of the order of 0.2 may be unacceptable. However, the corrector phase improves the predictions substantially and results in highly accurate distributions of displacements and stresses through the thickness. This is particularly true for the procedure based on calculating *a posteriori* the functional dependence of the displacements on the thickness coordinate. This procedure works well even for thick laminates with thickness-to-wavelength ratio of the order of 0.5.

ACKNOWLEDGMENT

The present research is partially supported by a NASA Grant No. NAG1-788 and by an Air Force Office of Scientific Research Grant No. 87-0115. The authors acknowledge useful discussions with James H. Starnes, Jr. of NASA Langley Research Center.

REFERENCES

1. Lekhnitski, S. G., *Anisotropic plates*, Fizmatgiz, Moscow, first edition 1947, second edition 1957 (in Russian); English translation published by Gordon and Breach, New York, 1968.
2. Ambartsumian, S. A., *Theory of anisotropic shells*, Fizmatgiz, Moscow, 1961 (in Russian); English translation NASA TT-F-118, 1964.
3. Ambartsumian, S. A., *General theory of anisotropic shells*, Izdatel'stvo Nauka, Moscow, 1974 (in Russian).
4. Bolotin, V. V. and Novichkov, Yu. N., *Mechanics of multilayered structures*, Izdatel'stvo Mashinostroenie, Moscow, 1980 (in Russian).
5. Grigorenko, Ya. M. and Vasilenko, A. T., *Theory of Shells with Variable Stiffness*, Methods of Calculation of Shells, Vol. 4, Izdatel'stvo Naukova Dumka, Kiev (in Russian), 1981.
6. Pelekh, B. L. and Lazko, V. A., *Layered Anisotropic Plates and Shells with Stress Concentration*, Izdatel'stvo Naukova Dumka, Kiev (in Russian), 1982.
7. Alfutov, N. A., Zinovev, P. A. and Popov, B. G., *Analysis of Multilayer Plates and Shells of Composite Materials*, Izdatel'stvo Mashinostroenie, Moscow, 1984 (in Russian).

8. Kovarik, V., *Stresses in Layered Shells of Revolution*, Prague, 1985; English translation published by Elsevier, NY, 1989.

9. Rasskazov, A. O., Sokolovskaya, I. I. and Shul'ga, N. A., *Theory and Analysis of Layered Orthotropic Plates and Shells*, Izdatel'stvo Obedinenia Vishcha Shkola, Kiev, 1986 (in Russian).

10. Ambartsumian, S. A., *Theory of anısotropic plates*, Nauka, Moscow, 1987 (in Russian).

11. Grigorenko, Ya. M., Vasilenko, A. T. and Golub, G. P., *Statics of Anisotropic Shells with Finite Shear Rigidity*, Izdatel'stvo Naukova, Dumka, Kiev (in Russian), 1987.

12. Vanin, G. A. and Semeniuk, N. P., *Stability of Shells of Composite Materials with Imperfections*, Izdatel'stvo Naukova Dumka, Kiev (in Russian), 1987.

13. Whitney, J. M., *Structural analysis of laminated anisotropic plates*, Technomic, Lancaster, PA, 1987.

14. Grigolyuk, E. I. and Kulikov, G. M., Multilayered reinforced shells - analysis of pneumatic tires, Izdatel'stvo *Mashinostroienie*, Moscow, 1988 (in Russian).

15. Grigorenko, Ya. M. and Kriukov, N. N., *Numerical Solutions of Static Problems for Elastic Layered Shells with Variable Parameters*, Izdatel'stvo Naukova Dumka, Kiev (in Russian), 1988.

16. Khoroshun, P. L., Kozlov, S. V., Ivanov, Yu. A. and Koshevoi, I. K., *Generalized Theory of Nonhomogeneous, in the Thickness Direction, Plates and Shells*, Izdatel'stvo Naukova Dumka, Kiev, 1988 (in Russian).

17. Vasilev, V. V., *Mechanics of Structures Made of Composite Materials*, Moscow, Izdatel'stvo Mashinostroenie, Moscow (in Russian), 1988.

18. Ambartsumian, S. A., Some current aspects of the theory of anisotropic layered shells, in *Applied Mechanics Surveys*, edited by H. N. Abramson, et al., Spartan Books, Washington, D.C., 1966, pp. 301-314.

19. Grigolyuk, E. I. and Kogan, F. A., State-of-the-art of the theory of multilayer shells, *Prikladnaya Mechanika*, 8(6), June 1972, 3-17 [English translation in *Soviet Applied Mechanics*, 8(6), July 1974, 583-595].

20. Bert, C. W., Analysis of Plates, and Analysis of shells, in *Composite Materials - Structural*

Design and Analysis. Part I, Vol. 7, edited by C. C. Chamis, Academic Press, NY, 1975, Chapter 4, 149-206 and Chapter 5, 207-258.

21. Grigolyuk, E. I. and Kulikov, G. M., General direction of development of the theory of multilayered shells, *Mekhanika Kompozitnykh Materialov*, **24**(2), March-April 1988, 287-298 [English translation in *Mechanics of Composite Materials*, **24(2)**, Sept. 1988, 231-241].

22. Noor, A. K. and Burton, W. S., Assessment of shear deformation theories for multilayered composite plates, *Applied Mechanics Reviews*, **42**(1), 1989, 1-13.

23. Kapania, R. K., A review on the analysis of laminated shells, *Journal of Pressure Vessel Technology*, ASME, **111**, May 1989, pp. 88-96.

24. Noor, A. K. and Burton, W. S., Assessment of computational models for multilayered composite shells, *Applied Mechanics Reviews* (to appear).

25. Dong, S. B., Pister, K. S. and Taylor, R. L., On the theory of laminated anisotropic shells and plates, *Journal of Aerospace Sciences*, **29(8)**, 1962, 969-975.

26. Pagano, N. J., Exact solutions for composite laminates in cylindrical bending, *Journal of Composite Materials*, **3**, 1969, 398-411.

27. Pagano, N. J., Exact solutions for rectangular bidirectional composites and sandwich plates, *Journal of Composite Materials*, **4**, 1970, 20-34.

28. Srinivas, S. and Rao, A. K. Bending, vibration and buckling of simply supported thick orthotropic rectangular plates and laminates, *International Journal of Solids and Structures*, **6**, 1970, 1463-1481.

29. Srinivas, S., Joga Rao, C. V. and Rao, A. K., An exact analysis for vibration of simply supported homogeneous and laminated thick rectangular plates, *Journal of Sound and Vibration*, **12(2)**, 1970, 187-199.

30. Grigorenko, Ya. M., Vasilenko, A. T. and Pankratova, N. D., Computation of the stressed state of thick-walled inhomogeneous anisotropic shells, *Prikladnaya Mekhanika*, **10(5)**, May 1974, 86-93 [English translation in Soviet Applied Mechanics, **10(5)**, Nov. 1975, 523-528].

31. Grigorenko, Ya. M., Vasilenko, A. T. and Pankratova, N. D., Stress state of composite

shells in the three-dimensional statement, *Mekhanika Kompozitnykh Materialov*, **20(4)**, July-Aug. 1984, 667-674 |English translation in *Mechanics of Composite Materials*, **20**(4), Jan. 1985, 468-474|.

32. Noor, A. K. and Peters, J. M., Stress, vibration an i buckling of multilayered cylinders, *Journal of Structural Engineering*, ASCE, **115(1)**, Jan. 1989, 69-88.

33. Noor, A. K. and Burton, W. S., Stress and free vibration analyses of multilayered composite plates, *Composite Structures*, **11**, 1989, 183-204.

34. Yang, P. C., Norris, C. H. and Stavsky, Y., Elastic wave propagation in heterogeneous plates, *International Journal of Solids and Structures*, **2**, 1966, 665-684.

35. Whitney, J. M., Stress analysis of thick laminated composite and sandwich plates,*Journal of Composite Materials*, **6**, 1972, 426-440.

36. Dong, S. B. and Tso, F. K. W., On a laminated orthotropic shell theory including transverse shear deformation, *Journal of Applied Mechanics*, ASME, **39**, Dec. 1972, 1091-1096.

37. Greenberg, J. B. and Stavsky, Y., Vibrations of axially compressed laminated orthotropic cylindrical shells, including transverse shear deformation, *Acta Mechanica*, **37(1-2)**, 1980, 13-28.

38. Reddy, J. N., Exact solutions of moderately thick laminated shells, *Journal of Engineering Mechanics*, ASCE, **110**(5), May 1984, 794-809.

39. Srinivas, S., A refined analysis of composite laminates, *Journal of Sound and Vibration*, **30(4)**, 1973, 495-507.

40. Sun, C. T. and Whitney, J. M., Theories for the dynamic response of laminated plates, *AIAA Journal*, **11(2)**, 1973, 178-183.

41. Lo, K. H., Christensen, R. M. and Wu, E. M., A high-order theory of plate deformation, Part 2: laminated plates, *Journal of Applied Mechanics*, ASME, **44(4)**, 1977, 669-676.

42. Levinson, M., An accurate, simple theory of the statics and dynamics of elastic plates, *Mechanics Research Communications*, **7(6)**, 1980, 343-350.

43. Murthy, M. V. V., *An Improved Transverse Shear Deformation Theory for Laminated Anisotropic Plates*, NASA Technical Paper 1903, Nov. 1981, 1-37.

44. Reddy, J. N., A simple higher-order theory for laminated composite plates, *Journal of*

Applied Mechanics, ASME, **51**, 1984, 745-752.

45. Phan, N. D. and Reddy, J. N., Analysis of laminated composite plates using a higher-order shear deformation theory, *International Journal for Numerical Methods in Engineering*, **21**, 1985, 2201-2219.

46. KrishnaMurthy, A. V. and Reddy, T. S. R., A higher-order theory of laminated composite cylindrical shells, *Journal of the Aeronautical Society of India*, **38**, Aug. 1986, 161-171.

47. Pandya, B. N. and Kant, T., A consistent refined theory for flexure of a symmetric laminate, *Mechanics Research Communications*, **14**(2), 1987, 107-113.

48. Noor, A. K., Stability of multilayered composite plates, *Fibre Science and Technology*, **8**, April 1975, 81-89.

49. Noor, A. K. and Peters, J. M., A posteriori estimates for shear correction factors in multilayered composite cylinders, *Journal of Engineering Mechanics*, ASCE, **115**(6), June 1989, 1225-1244.

50. Noor, A. K. and Burton, W. S., Three-Dimensional Solutions for Antisymmetrically Laminated Anisotropic Plates, *Journal of Applied Mechanics* (to appear).

51. Noor, A. K. and Burton, W. S., Assessment of Computational Models for Multilayered Anisotropic Plates, *Composite Structures* (to appear).

52. Lekhnitskii, S. G., *Theory of Elasticity of an Anisotropic Body*, Mir Publishers, Moscow, 1981.

53. Hearmon, R. F. S., *An Introduction to Applied Anisotropic Elasticity*, Oxford University Press, London, 1961.

54. Reissner, E., On a certain mixed variational theorem and laminated shell theory, in *Refined Dynamical Theory of Beams, Plates and Shells* (Proceedings of Euromechanics Colloquium, No. 219), Springer-Verlag, 1987, 17-27.

55. Reissner, E., On a mixed variational theorem and on shear deformable plate theory, *International Journal for Numerical Methods in Engineering*, **23**, 1986, 193-198.

56. Srinivas, S., Analysis of laminated, composite, circular cylindrical shells with general boundary conditions, NASA TR-R-412, April 1974.

57. Whitney, J. M., Shear correction factors for orthotropic laminates under static load, *Journal of Applied Mechanics*, ASME, **40**(1), 1973, 302-304.

58. Chow, T. S., On the propagation of flexural waves in an orthotropic laminated plate and its response to an impulsive load, *Journal of Composite Materials*, **5**, 1971, 306-319.

Table 1 - Modeling Approaches Used in the Numerical Studies

Model No.	Description	Through-the-Thickness Displacement Assumptions	Constraint Conditions on Stresses	Total Number of Displacement Parameters
1, 1A	First-order shear deformation theory	• linear u_1,u_2 • constant w	$\sigma_{33}=0$	5
2	First-order theory with transverse normal stresses and strains included	linear u_1,u_2 and w	none	6
3	Higher-order shear deformation theory	quintic u_1,u_2 and w	none	18
4	Simplified higher-order theory	• cubic u_1,u_2 • constant w	$\sigma_{33}=0$ throughout and σ_{13} and $\sigma_{23}=0$ at top and bottom surfaces	5
5	Discrete layer theory	• piecewise linear u_1,u_2 • constant w (through-the-thickness)	$\sigma_{33}=0$ throughout	2×NL+3
6, 6A	Predictor-corrector procedures	Predictor Phase • linear u_1,u_2 • constant w Corrector Phase See Note 2	Predictor Phase $\sigma_{33}=0$ Corrector Phase None	5

Notes: 1) In model 1, $k_1=k_2=1$, and in model 1A, they are computed from the cylindrical bending condition of Refs. 57 and 58.

2) In model 6, the corrector phase is based on adjusting the transverse shear stiffnesses (see Refs. 33 and 49) and in model 6A it is based on correcting the thickness distribution of the in-plane and transverse displacement components.

Table 2 - Effect of thickness ratio, h/L_1, and fiber orientation angle θ, on the accuracy of the fundamental vibration frequencies obtained by predictor-corrector procedures and other modeling approaches (see Table 1). Antisymmetrically laminated composite plates with NL=10.

h/L_1	Ω_{exact}	$U_2/U\times10^2$	Values of $\omega^2/\omega^2_{exact}$						
			Model 1	Model 1A	Model 3	Model 4	Model 5	Model 6	Model 6A
					a) θ=15°				
0.1	1.348	22.57	1.041	1.003	1.003	1.003	1.000	0.9967	1.000
0.2	12.82	51.09	1.097	1.003	1.008	1.008	1.001	0.9878	1.000
0.3	39.51	66.66	1.125	0.9988	1.010	1.016	1.003	0.9764	1.002
0.4	81.93	74.65	1.136	0.9921	1.012	1.027	1.005	0.9648	1.004
					b) θ=30°				
0.1	1.670	26.02	1.057	1.013	1.010	1.012	1.002	0.9967	1.000
0.2	14.95	56.98	1.131	1.026	1.022	1.030	1.007	0.9884	1.001
0.3	44.11	73.17	1.170	1.027	1.028	1.047	1.012	0.9780	1.003
0.4	88.68	81.07	1.185	1.023	1.032	1.066	1.017	0.9675	1 007
					c) θ=45°				
0.1	1.813	28.02	1.065	1.018	1.014	1.017	1.003	0.9966	1.000
0.2	15.71	59.61	1.146	1.035	1.029	1.040	1.009	0.9881	1.001
0.3	45.52	75.44	1.187	1.037	1.037	1.060	1.015	0.9775	1.003
0.4	90.57	82.96	1.203	1.033	1.041	1.083	1.021	0.9666	1.009

Notes: 1) $\Omega_{exact} = 10^2 \times \rho h^2\ \omega^2_{exact}/E_T$. In model 1, $k_1=k_2=1$, and in model 1A, k_1 and k_2=5/6 as computed from the cylindrical bending condition of Refs. 57 and 58.

2) U_2 is the transverse shear strain energy (associated with $\sigma_{\alpha3}$ and $2\varepsilon_{\alpha3}$).

Table 3 - Maximum absolute values of displacements, stresses and transverse shear strain energy density obtained by the three-dimensional elasticity model. Antisymmetrically laminated composite plates subjected to static loading $p=p_o \sin \pi\xi_1 \sin \pi\xi_2$, $h/L_1 = 0.3$, NL=10 and $\theta=45°$ (see Fig. 5).

Quantity	Maximum Absolute Value
$u_{1,A} E_T/p_o h$	.354
$w_S E_T/p_o h$	2.19
$\sigma_{11,A}/p_o$	2.48
$\sigma_{13,S}/p_o$	.730
$\sigma_{33,A}/p_o$	.500
$U_{13,S} E_T/(p_o)^2$	.662

Note: Subscripts S and A refer to the symmetric and antisymmetric components (in the thickness coordinate x_3).

Table 4 - Effect of thickness ratio, h/r_o, and circumferential wave number, n, on the accuracy of the lowest vibration frequencies obtained by predictor-corrector procedures and other modeling approaches. Simply-supported composite cylinders with NL=10 (see Table 1).

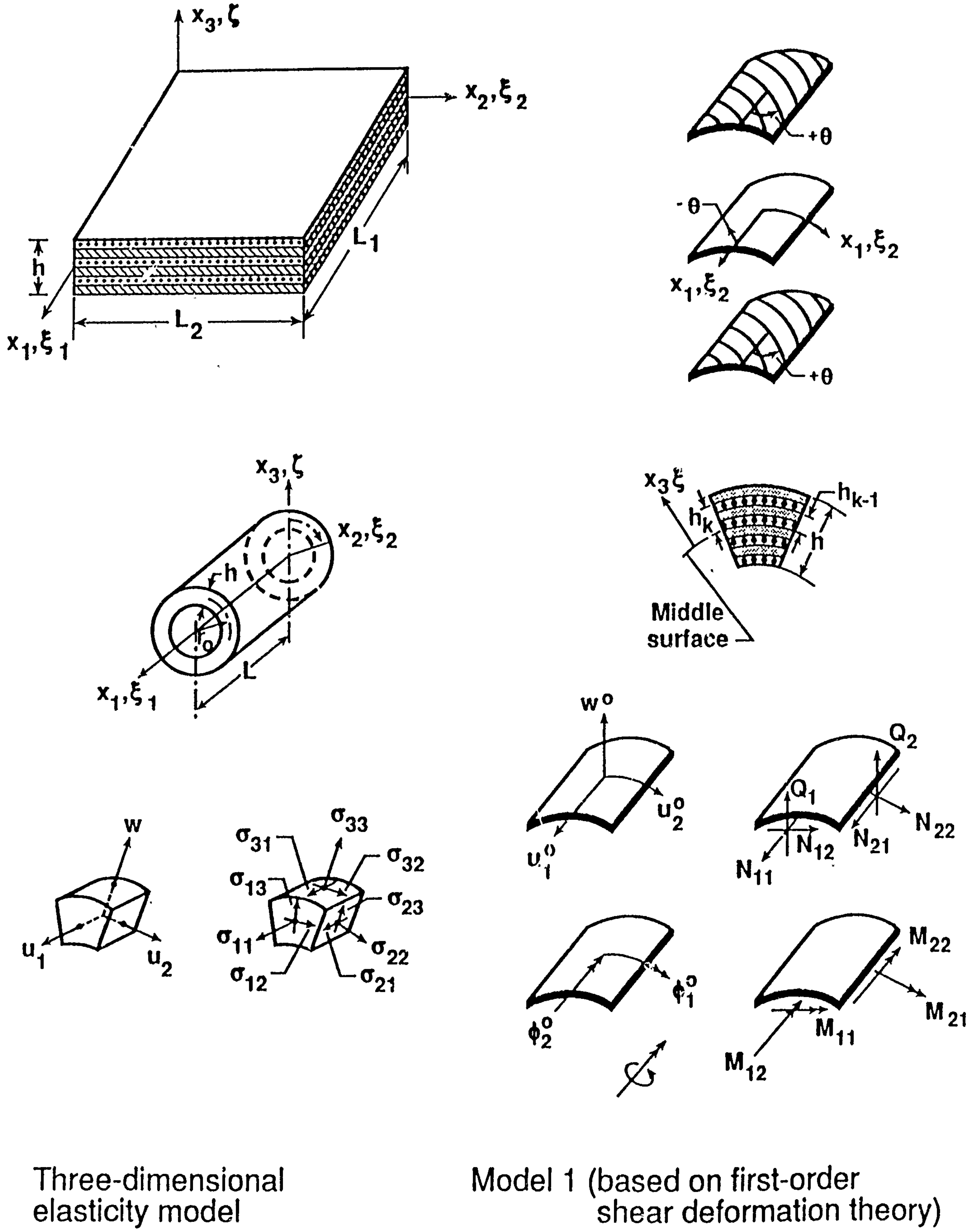
n	Ω_{exact}	$U_2/U \times 10^2$	Values of $\omega^2/\omega^2_{exact}$						
			Model 1	Model 1A	Model 3	Model 4	Model 5	Model 6	Model 6A
					a) h/r_o=0.05				
0	1.234	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1	.5447	.1950	1.001	1.000	1.000	1.000	1.000	1.000	1.000
2	.2557	.6523	1.002	1.001	1.000	1.001	1.001	1.001	1.000
3	.1753	1.898	1.006	1.002	1.001	1.003	1.003	1.002	1.000
4	.1938	4.910	1.014	1.003	1.002	1.006	1.003	1.003	1.000
5	.2987	9.266	1.024	1.002	1.004	1.008	1.003	1.003	1.000
6	.5036	13.95	1.035	1.002	1.007	1.010	1.002	1.002	1.000
7	.8288	18.65	1.045	1.001	1.009	1.012	1.002	1.001	1.000
8	1.295	23.30	1.056	1.000	1.011	1.015	1.001	1.000	1.000
9	1.920	27.87	1.067	.9992	1.013	1.017	1.001	.9985	1.000
10	2.718	32.29	1.078	.9984	1.016	1.020	1.001	.9975	1.000
					b) h/r_o=0.3				
0	44.41	0	1.009	1.009	1.000	1.000	1.000	1.009	1.000
1	29.88	22.02	1.057	.9991	1.011	1.016	1.006	.9988	1.002
2	25.69	41.70	1.106	.9975	1.020	1.031	1.015	.9983	1.003
3	33.10	55.42	1.137	.9913	1.027	1.039	1.016	.9898	1.004
4	49.62	66.14	1.162	.9844	1.032	1.043	1.014	.9770	1.005
5	74.10	73.78	1.180	.9775	1.036	1.049	1.012	.9645	1.007
6	106.0	78.93	1.190	.9704	1.038	1.056	1.010	.9533	1.010
7	144.9	82.39	1.194	.9632	1.039	1.066	1.010	.9433	1.016
8	190.9	84.74	1.195	.9562	1.040	1.079	1.010	.9342	1.024
9	244.0	86.37	1.194	.9493	1.040	1.095	1.010	.9260	1.031
10	304.0	87.51	1.191	.9427	1.039	1.114	1.011	.9183	1.038

Notes: 1) $\Omega_{exact}=10^2 \times \rho h^2 \omega^2_{exact}/E_T$. In model 1, $k_1=k_2=1$, and in model 1A, $k_1=k_2=0.7731$, as computed from the cylindrical bending condition of Refs. 57 and 58.

2) U_2 is the transverse shear strain energy (associated with $\sigma_{\alpha3}$ and $2\varepsilon_{\alpha3}$).

Table 5 - Relative magnitudes of the maximum displacements and stresses obtained by the three-dimensional elasticity model. Simply supported composite cylinders subjected to internal normal pressure $p=p_o \sin \pi\xi_1 \cos 2\pi\xi_2$, $h/r_o=0.3$, NL=10 and $L/r_o=1.0$ (see Fig. 8).

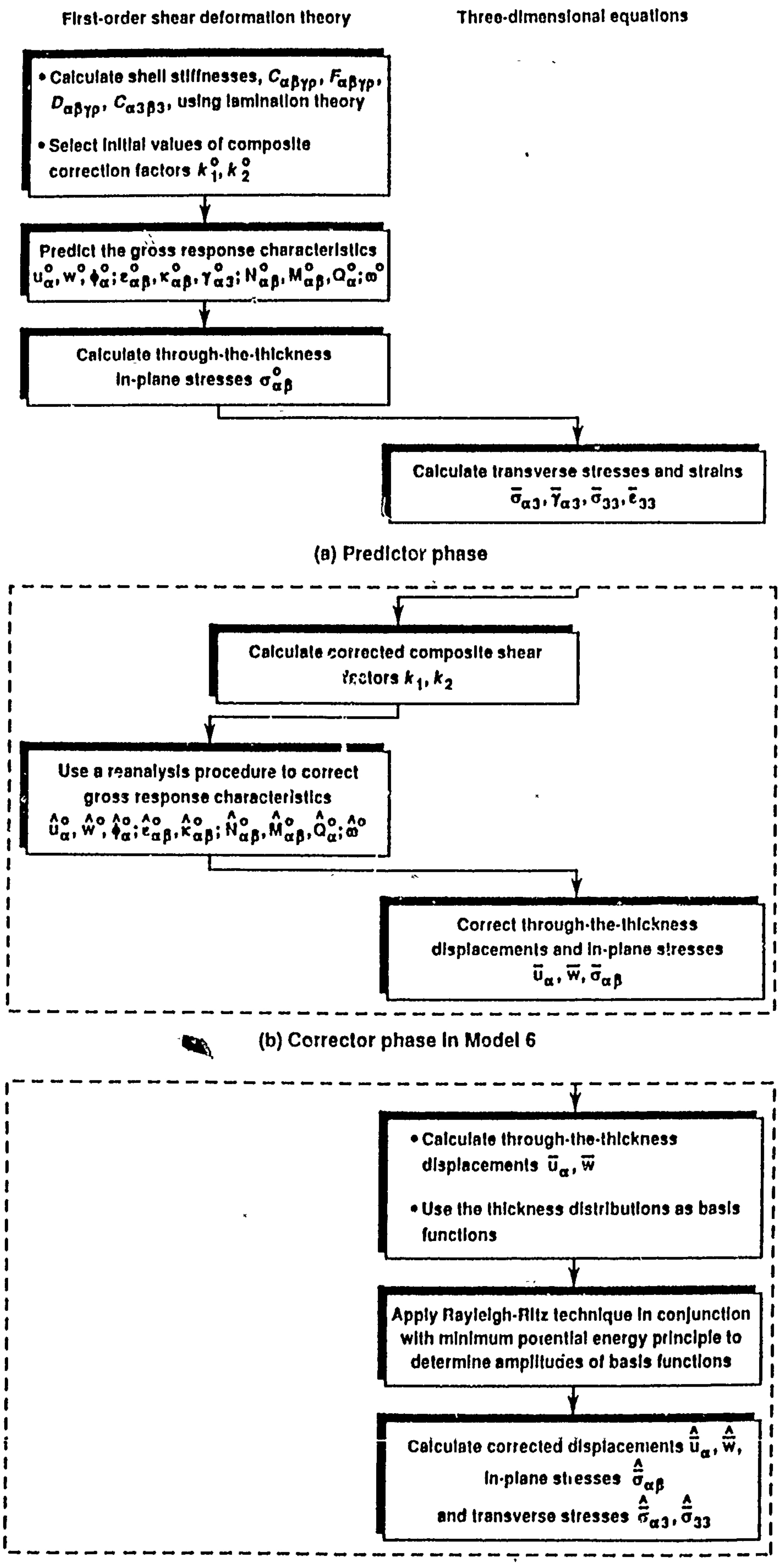
Quantity	Maximum Absolute Value
$u_1 E_T/p_o h$	0.489
$u_2 E_T/p_o h$	1.24
$w E_T/p_o h$	2.18
σ_{11}/p_o	5.18
σ_{12}/p_o	.643
σ_{13}/p_o	.566
$U_{13} E_T/(p_o)^2$	.476



Three-dimensional elasticity model

Model 1 (based on first-order shear deformation theory)

Figure 1


First-order shear deformation theory
Three-dimensional equations
• Calculate shell stiffnesses, $C_{\alpha\beta\gamma\rho}$, $F_{\alpha\beta\gamma\rho}$, $D_{\alpha\beta\gamma\rho}$, $C_{\alpha 3\beta 3}$, using lamination theory
• Select initial values of composite correction factors k_1^0, k_2^0
Predict the gross response characteristics $u_\alpha^0, w^0, \phi_\alpha^0; \varepsilon_{\alpha\beta}^0, \kappa_{\alpha\beta}^0, \gamma_{\alpha 3}^0; N_{\alpha\beta}^0, M_{\alpha\beta}^0, Q_\alpha^0; \omega^0$
Calculate through-the-thickness in-plane stresses $\sigma_{\alpha\beta}^0$
Calculate transverse stresses and strains $\bar{\sigma}_{\alpha 3}, \bar{\gamma}_{\alpha 3}, \bar{\sigma}_{33}, \bar{\varepsilon}_{33}$
Calculate corrected composite shear factors k_1, k_2
Use a reanalysis procedure to correct gross response characteristics $\hat{u}_\alpha^0, \hat{w}^0, \hat{\phi}_\alpha^0; \hat{\varepsilon}_{\alpha\beta}^0, \hat{\kappa}_{\alpha\beta}^0; \hat{N}_{\alpha\beta}^0, \hat{M}_{\alpha\beta}^0, \hat{Q}_\alpha^0; \hat{\omega}^0$
Correct through-the-thickness displacements and in-plane stresses $\bar{u}_\alpha, \bar{w}, \bar{\sigma}_{\alpha\beta}$
• Calculate through-the-thickness displacements $\bar{u}_\alpha, \bar{w}$
• Use the thickness distributions as basis functions
Apply Rayleigh-Ritz technique in conjunction with minimum potential energy principle to determine amplitudes of basis functions
Calculate corrected displacements $\hat{\bar{u}}_\alpha, \hat{\bar{w}}$, in-plane stresses $\hat{\bar{\sigma}}_{\alpha\beta}$ and transverse stresses $\hat{\bar{\sigma}}_{\alpha 3}, \hat{\bar{\sigma}}_{33}$

(a) Predictor phase

(b) Corrector phase in Model 6

(c) Corrector phase in Model 6A

Figure 2

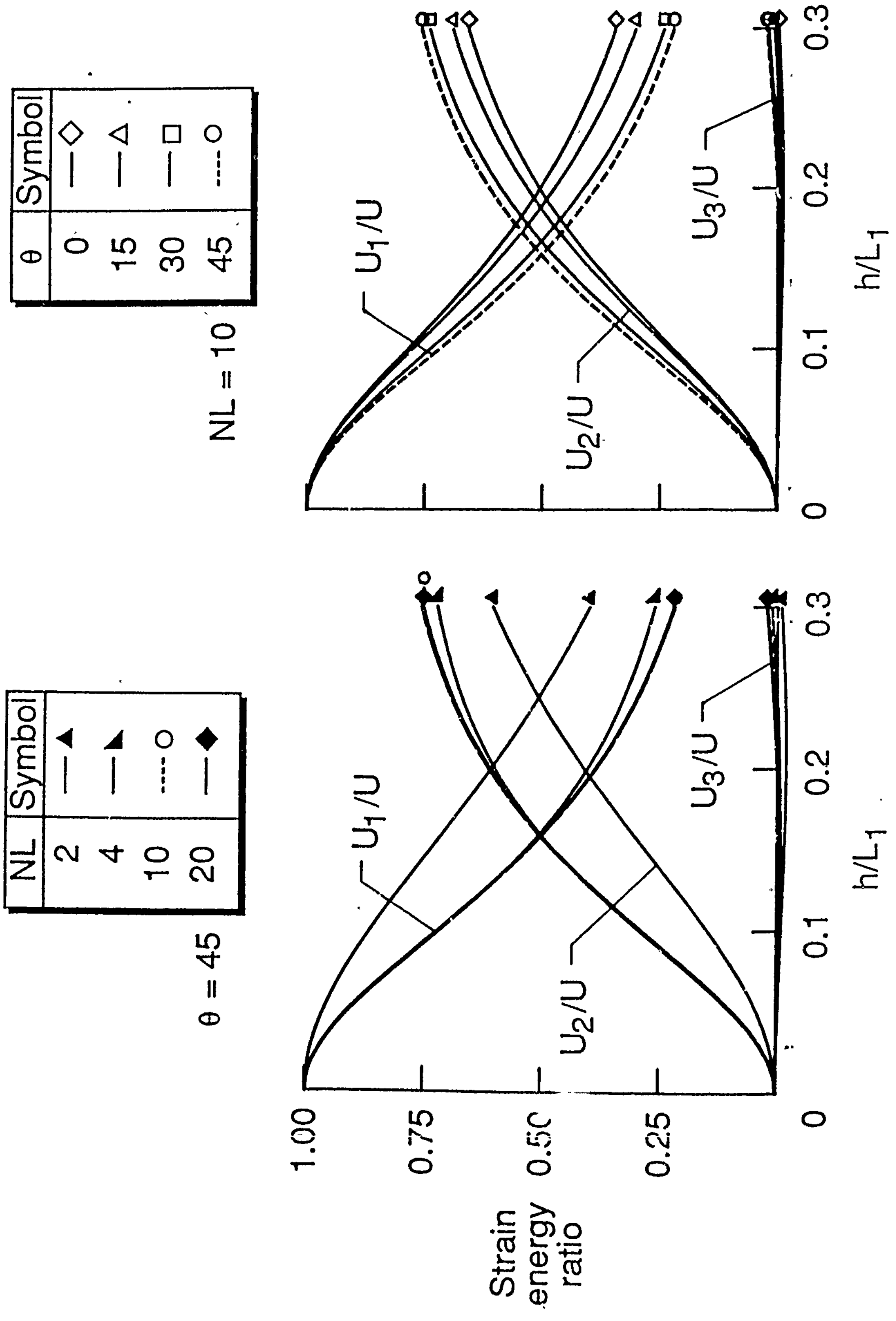

θ = 45
NL
Symbol
2
4
10
20
NL = 10
θ
Symbol
0
15
30
45
Strain energy ratio
1.00
0.75
0.50
0.25
0
0.1
0.2
0.3
h/L1
U1/U
U2/U
U3/U

Figure 3

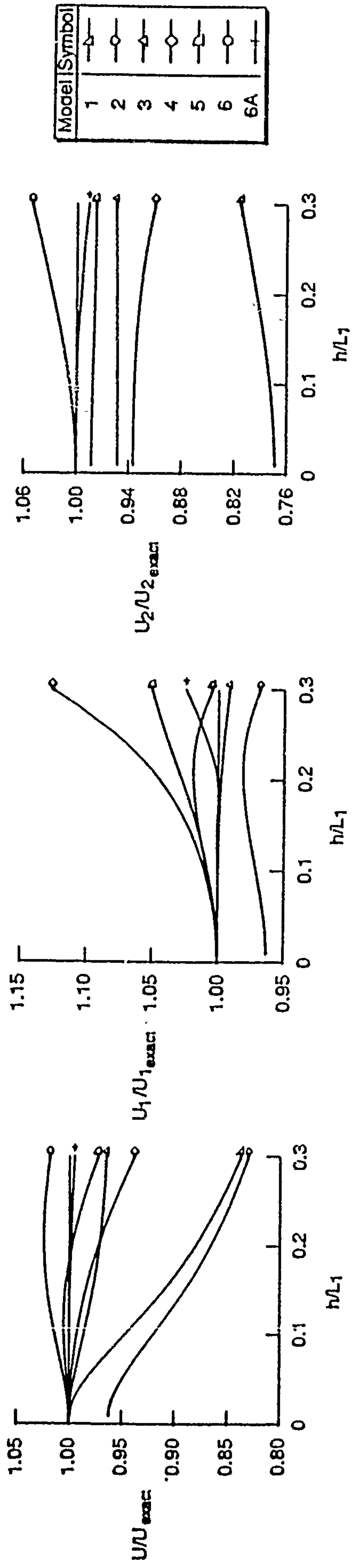
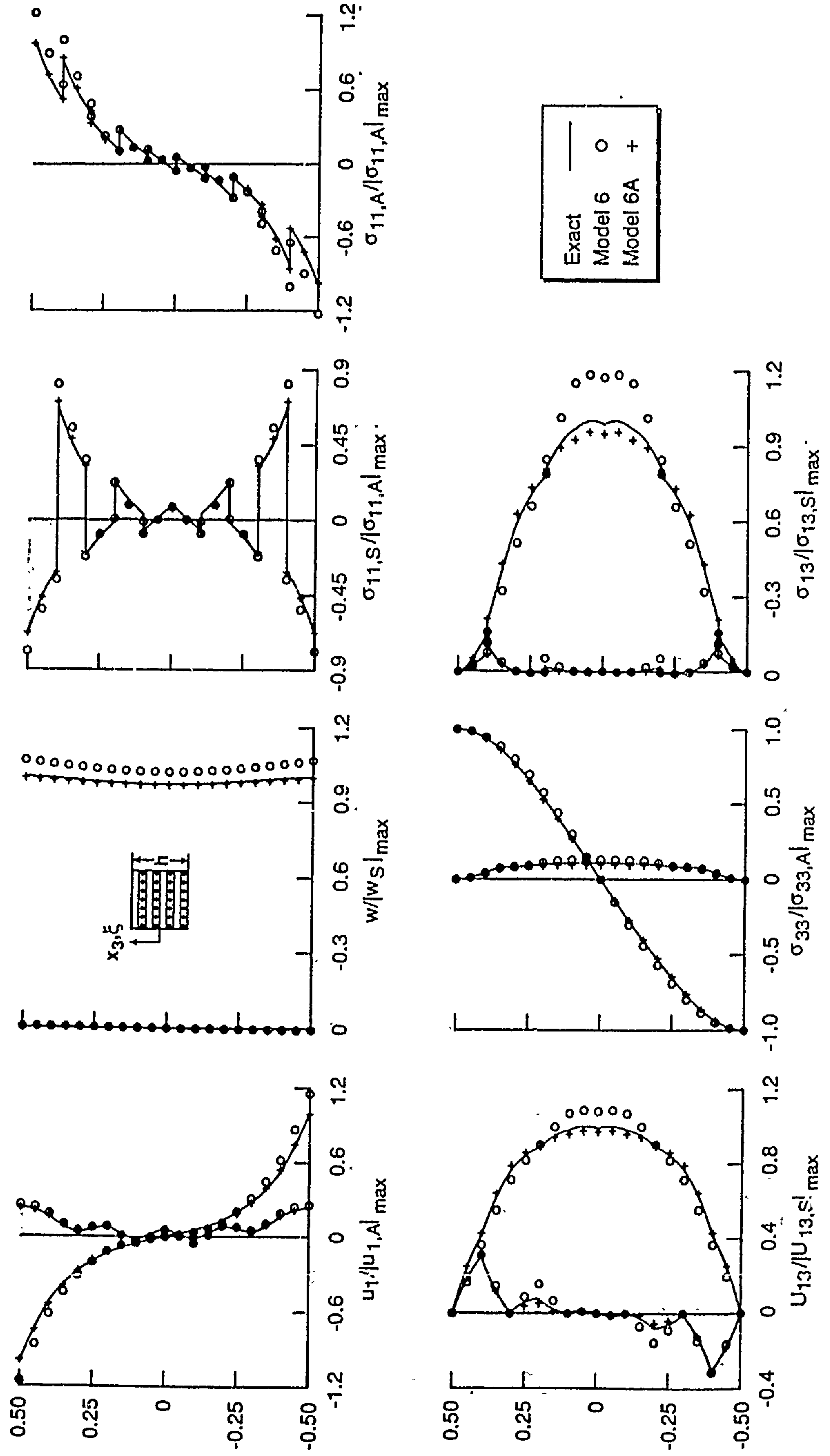

Model
Symbol
1
2
3
4
5
6
6A
U/U_exact
1.05
1.00
0.95
0.90
0.85
0.80
0
0.1
0.2
0.3
h/L_1
U_1/U_1exact
1.15
1.10
1.05
1.00
0.95
0
0.1
0.2
0.3
h/L_1
U_2/U_2exact
1.06
1.00
0.94
0.88
0.82
0.76
0
0.1
0.2
0.3
h/L_1

Figure 4



ζ
0.50
0.25
0
-0.25
-0.50
$u_1/|u_{1,A}|_{max}$
-1.2
-0.6
0
0.6
1.2
$w/|w_S|_{max}$
0
-0.3
0.6
0.9
1.2
x_3, ξ
h
$\sigma_{11,S}/|\sigma_{11,A}|_{max}$
-0.9
-0.45
0
0.45
0.9
$\sigma_{11,A}/|\sigma_{11,A}|_{max}$
-1.2
-0.6
0
0.6
1.2
$U_{13}/|U_{13,S}|_{max}$
-0.4
0
0.4
0.8
1.2
$\sigma_{33}/|\sigma_{33,A}|_{max}$
-1.0
-0.5
0
0.5
1.0
$\sigma_{13}/|\sigma_{13,S}|_{max}$
0
-0.3
0.6
0.9
1.2
Exact
Model 6 ○
Model 6A +

Figure 5

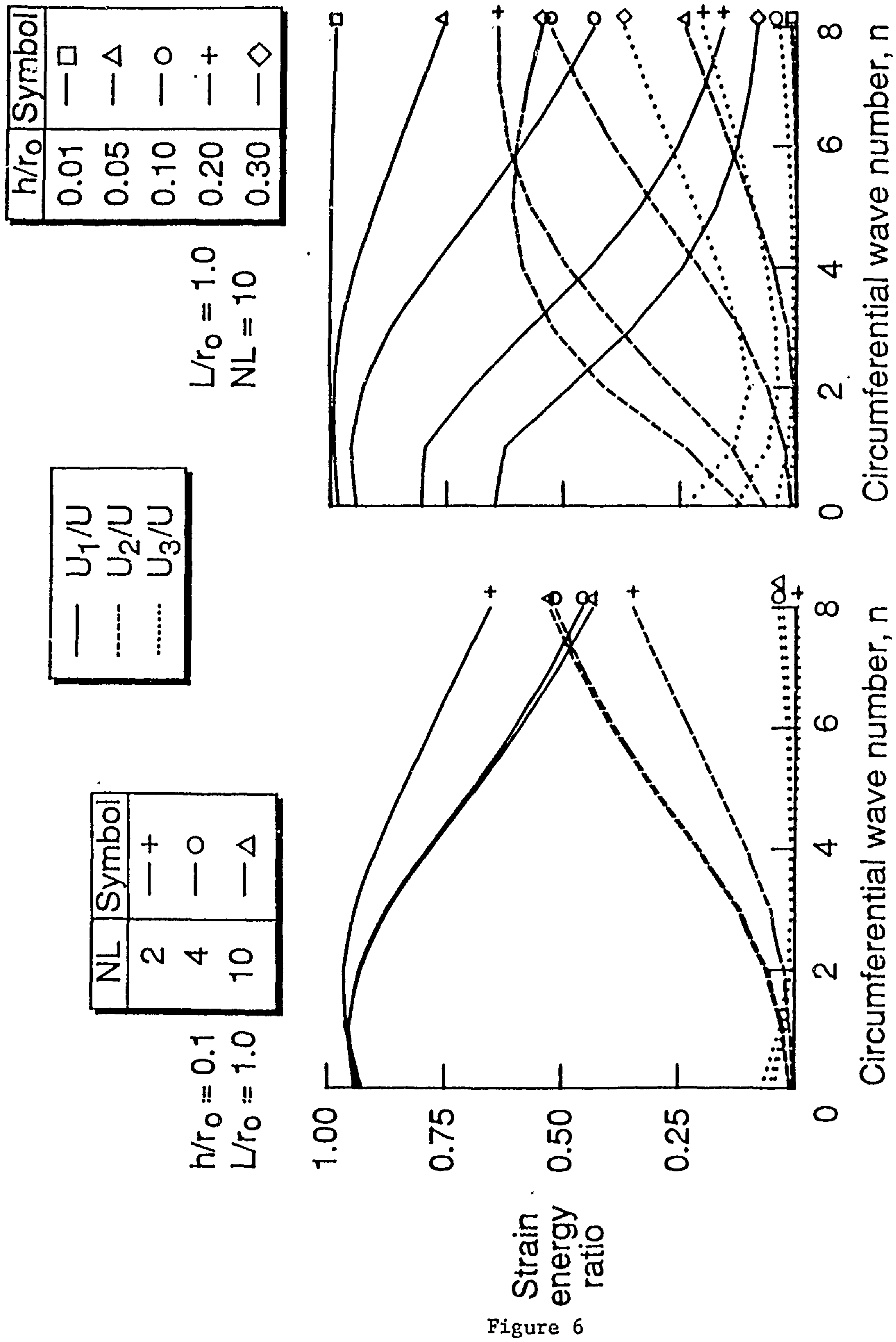

Strain energy ratio
1.00
0.75
0.50
0.25
0
2
4
6
8
Circumferential wave number, n
h/r_o = 0.1
L/r_o = 1.0
NL | Symbol
2 | —+
4 | —o
10 | —Δ
—— U_1/U
- - - U_2/U
······ U_3/U
L/r_o = 1.0
NL = 10
h/r_o | Symbol
0.01 | —□
0.05 | —Δ
0.10 | —o
0.20 | —+
0.30 | —◇

Figure 6

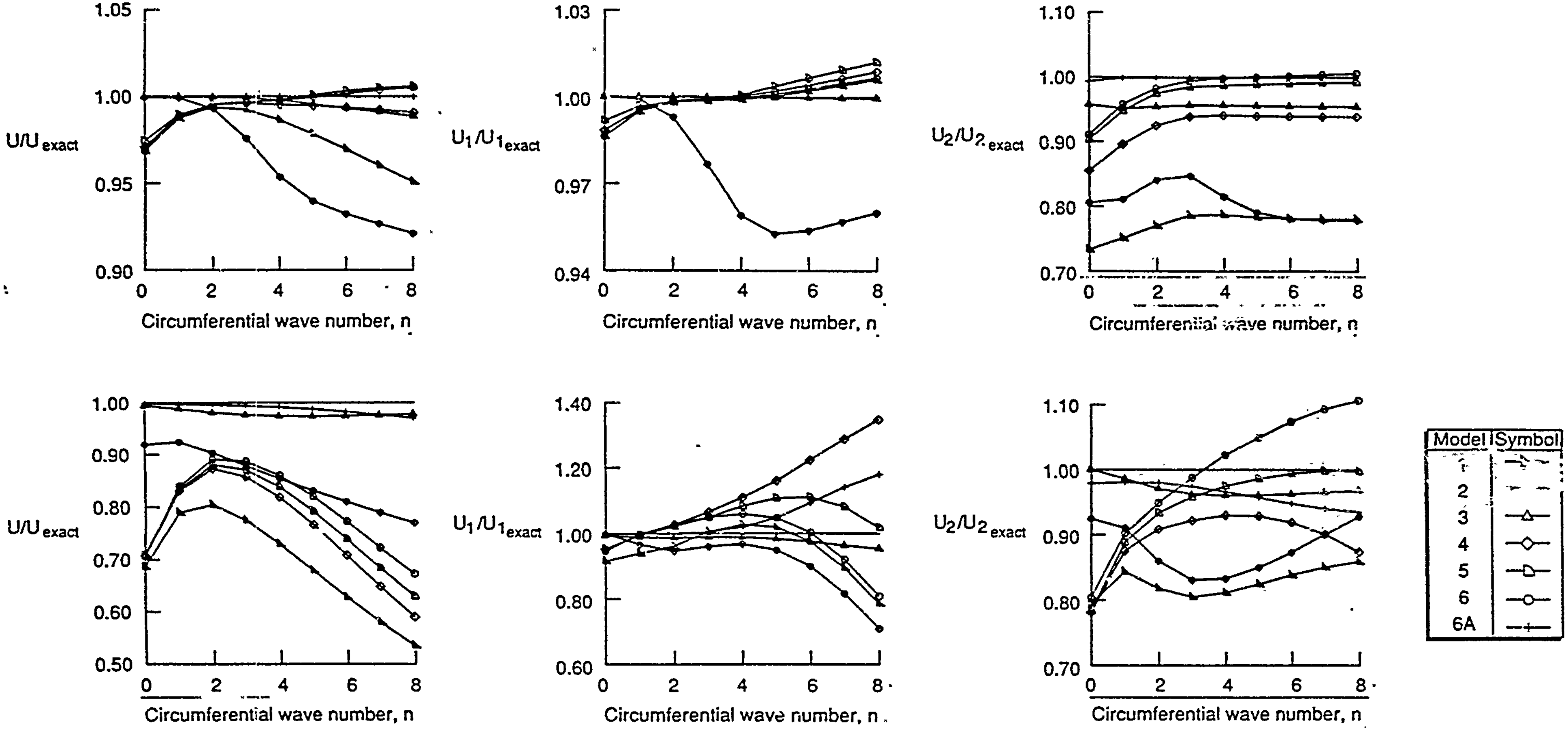

U/U_exact
U_1/U_1exact
U_2/U_2exact
1.05
1.00
0.95
0.90
1.03
0.97
0.94
1.10
0.80
0.70
0.60
0.50
1.40
1.20
0
2
4
6
8
Circumferential wave number, n
Model
Symbol
1
2
3
4
5
6
6A

Figure7

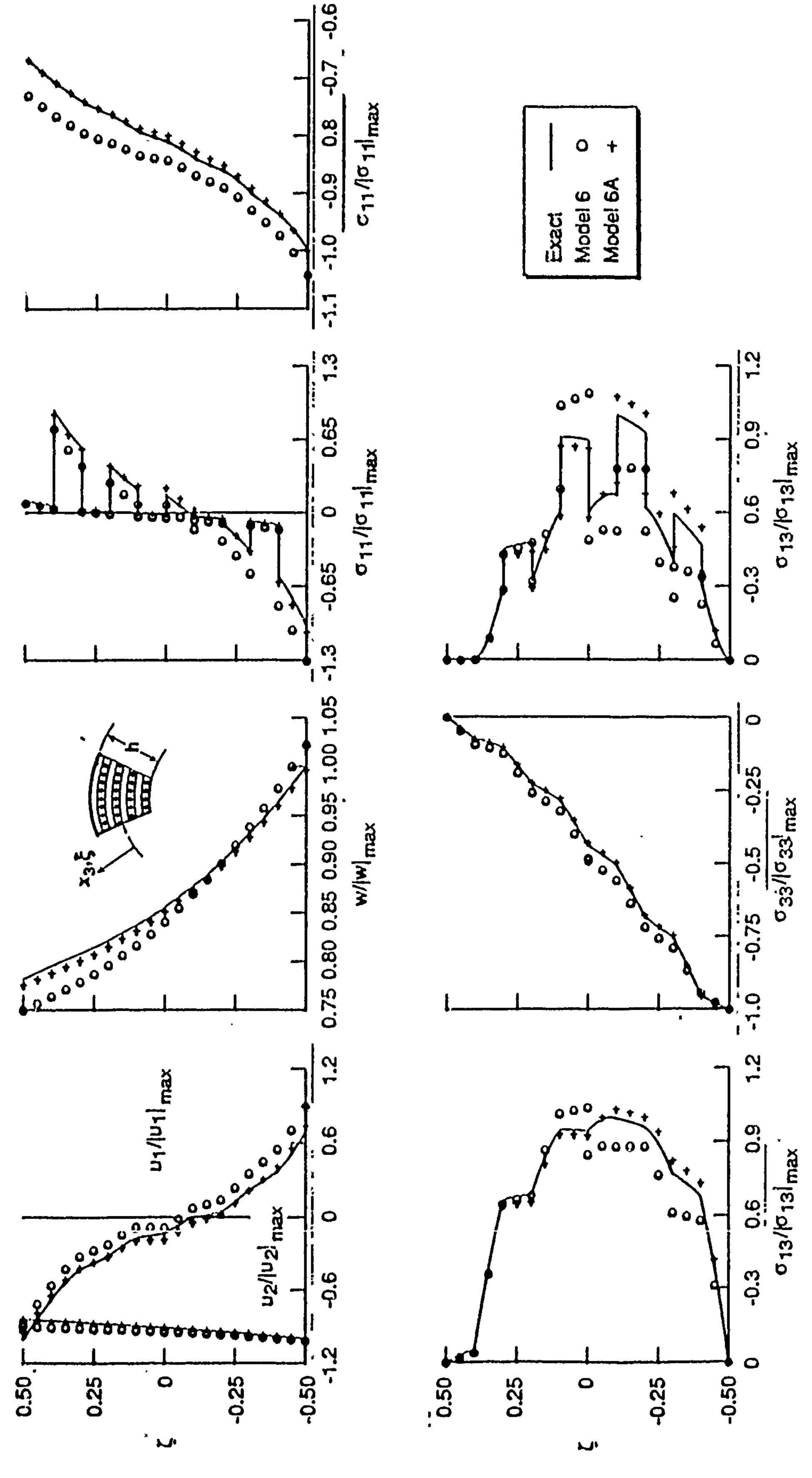

Exact
Model 6 o
Model 6A +
ζ
0.50
0.25
0
-0.25
-0.50
$u_1/|u_1|_{max}$
$u_2/|u_2|_{max}$
-1.2
-0.6
0
0.6
1.2
$w/|w|_{max}$
0.75 0.80 0.85 0.90 0.95 1.00 1.05
x_3, ξ
h
$\sigma_{11}/|\sigma_{11}|_{max}$
-1.3
-0.65
0
0.65
1.3
-1.1
-1.0
-0.9
0.8
-0.7
-0.6
$\sigma_{13}/|\sigma_{13}|_{max}$
0
0.3
0.6
0.9
1.2
$\sigma_{33}/|\sigma_{33}|_{max}$
-1.0
-0.75
-0.5
-0.25
0
-0.3
0.6
0.9
1.2

Figure 8

LIST OF TABLES

LIST OF FIGURES

Figure 8 - Accuracy of displacements, stresses and transverse shear strain energy density obtained by predictor-corrector procedures. Simply supported composite cylinder subjected to internal pressure $p=p_o \sin \pi\xi_1 \cos 2\pi\xi_2$, NL=10, and $h/r_o=0.3$. $U_{13} = \frac{1}{2}\sigma_{13} \times 2\varepsilon_{13}$. The normalization factors are given in Table 5.